NEW YORK UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES
LIBRARY
25 Waverly Place, New York 3, N. Y.

AFCRC-TN-56-562

EM 92

NEW YORK UNIVERSITY
Institute of Mathematical Sciences
Division of Electromagnetic Research

RESEARCH REPORT No. EM-92

Diffraction by an Aperture. I

JOSEPH B. KELLER

CONTRACT NO. AF 19(122)-42
JUNE, 1956

EM-92
c.1

NEW YORK UNIVERSITY
Institute of Mathematical Sciences
Division of Electromagnetic Research

Research Report No. EM-92

DIFFRACTION BY AN APERTURE. I

Joseph B. Keller

Joseph B. Keller
Joseph B. Keller

Morris Kline
Morris Kline
Project Director

June, 1956

The research reported in this document has been made possible through support and sponsorship extended by the Air Force Cambridge Research Center, under Contrcct No. AF 19(122)-42. It is published for technical information only and does not necessarily represent recommendations for conclusions of the sponsoring agency.

New York, 1956

Abstract

The diffraction of a wave by an aperture of any shape in a thin screen is treated by a new method. This method is called " the geometrical theory of diffraction", because it is an extension of geometrical optics which accounts for diffraction. In this method new rays - called diffracted rays - are introduced. They are produced when an incident ray hits the edge of the aperture, and they satisfy the " law of diffraction". A field is associated with each ray in a quantitative way, by means of the optical principles of phase variation and energy conservation. In addition " diffraction coefficients" are introduced to relate the field on a diffracted ray to that on the corresponding incident ray.

By this method a simple formula is obtained for the field diffracted by any aperture. By means of this formula the field in the aperture or the far field diffraction pattern can be found. In addition the transmission cross section of the aperture can be determined. Explicit formulas and numerical results are given for slits and circular apertures. The accuracy of the results increases as the wavelength decreases, but they are even useful for wavelengths as large as the aperture dimensions.

Table of Contents

Page

1. Introduction 1

2. Ray tracing 4

3. The field on a ray 6

4. Reflection and diffraction coefficients 9

5. The total field 15

6. The diffraction of spherical and cylindrical waves by a half-plane 19

7. Diffraction by a slit 21

8. Diffraction by a circular aperture 28

9. Diffraction by a convex aperture 36

10. Electromagnetic diffraction by a convex aperture 43

11. Diffraction by a grating of slits 46

12. Conclusion 48

Appendix I Caustics of diffracted rays 50

II The diffraction coefficient D 52

III The electromagnetic diffraction matrix 54

IV The field on an axial caustic 56

V Caustics of doubly diffracted rays 58

References 60

1. Introduction

The present treatment of diffraction of a wave by an aperture in a thin screen is based upon a new theory of wave propagation. This theory is called 'the geometrical theory of diffraction'[1], because it is an extension of geometrical optics which describes diffraction phenomena. It is like geometrical optics inasmuch as it postulates that fields propagate along certain lines or curves called rays. However, it goes beyond geometrical optics, by introducing various new kinds of rays, called 'diffracted rays', in addition to the usual optical ones. It also assigns a field value to each point on a ray in a quantitative way and postulates that the total field at a point is the sum of the fields on all rays through that point. Thus the new theory not only provides a qualitative description of diffraction in terms of the diffracted rays, but it permits a quantitative determination of the field as well.

By using this theory we have obtained a relatively simple formula for the field diffracted by an aperture of any shape. This formula takes into account the nature of the field under consideration, i.e. whether it is electromagnetic, acoustic or of some other type, and the nature of the screen. It enables us to compute the field in the aperture as well as in front of and behind the screen. In particular it yields the diffraction pattern and the transmission cross section of the aperture. The formula also applies to diffraction by a flat thin obstacle of any shape, and thus permits a verification of Babinet's principle. It also accounts for the bright lines and bright spots which have been observed in the shadows of such obstacles.

Because our theory is related to geometrical optics, our results should be valid when the wavelength λ is small compared to the aperture dimensions. We have verified this expectation by comparing our results with certain previously known exact results. In doing so we have also found that our results are accurate enough to be of use even for wavelengths as large as the aperture dimensions. Of course accuracy increases as the wavelength decreases.

Previously we have used the 'geometrical theory of diffraction' to treat diffraction by a cylinder of any cross section[2], diffraction by a smooth three-dimensional object of any shape[3], and other diffraction problems[4]. In these cases we also found excellent agreement between our results and previously known exact results. In fact, whenever comparison was possible, our results agreed perfectly with the leading terms in the asymptotic expansions of the exact solutions of diffraction problems, the expansions being with respect to $k = 2\pi/\lambda$, as $k \to \infty$. This agreement has confirmed our belief, which is based upon various mathematical considerations, that our theory always yields the leading terms in the asymptotic expansion of the exact solution. Since such expansions are often accurate even for small values of k (large values of λ), the accuracy of our results for large wavelengths is not surprising.

Certain aspects of our theory are, like geometrical optics, the same for many kinds of fields. This is true of the rules governing the rays, both ordinary and diffracted, and the principles determining the variation of phase and amplitude along a ray. Therefore these parts of the theory can be presented in a general manner, without reference to the nature of any particular field. On the other hand there are other aspects of the theory which depend essentially upon the nature of the field. This is the case with coefficients of reflection and transmission, and with the diffraction coefficients which we will introduce. Consequently these coefficients must be considered separately for different kinds of fields. Because of this we will first present the theory for a scalar field (e.g. acoustic pressure) and later show what differences occur in the treatment of electromagnetic fields.

We will formulate our problem as follows: A thin screen containing an aperture lies in the plane $x = 0$. The rest of space is filled with an homogeneous medium with propagation speed c. A given wave, with wavelength λ and propagation constant $k = \frac{2\pi}{\lambda}$, is incident upon the screen from the left. We wish to find the field which results from the interaction between the incident wave and the screen.

The incident wave can be characterized in either of two ways. The first characterization consists in giving the set of incident rays and the field value (i.e. phase and amplitude) on each ray. The second alternative consists in giving the value of the incident field u_{inc} at each point in the form

$$u_{inc}(x,y,z) = A(x,y,z)\, e^{ik\Psi(x,y,z)} . \tag{1}$$

From the second form (1) the rays can be determined as the orthogonal trajectories of the wavefronts $\Psi(x,y,z)$ = constant. Conversely, when the rays and the phase on each ray are known, $\Psi(x,y,z)$ can be found. Thus the two characterizations are equivalent. Therefore we will assume that $u_{inc}(x,y,z)$ is given in the form (1) and that the incident rays are also known.

In Section 2 we will determine the rays--especially the diffracted ones--which are produced when the incident rays hit the screen. In Section 3 we will determine how the field varies along a ray. In Section 4 we will determine the initial value of the field on each ray. Then in Section 5 we will obtain the formula for the field at any point by adding together the fields on all rays through the point. In subsequent sections the results will be applied to diffraction by a half-plane, a slit, a circular aperture, a convex aperture, and a grating of slits.

Most of the previous work on diffraction by an aperture has been carefully discussed by C.J. Bouwkamp[16].

2. Ray tracing

According to geometrical optics, the only rays which occur in the present problem are the incident rays and the reflected rays to which they give rise when they hit the screen. All of them are straight lines, and the direction of each reflected ray is determined by the law of reflection.

We assume, however, that there are additional rays which are produced when incident rays hit the edge of the screen. We call them 'diffracted rays' and assume that they are straight lines and that they satisfy the 'law of edge diffraction.' This law states that a diffracted ray and the corresponding incident ray make equal angles with the edge and lie on opposite sides of the plane normal to the edge at the point of diffraction. From this law it follows that there are infinitely many diffracted rays produced by one incident ray. These diffracted rays lie on the surface of a cone having as axis the tangent to the edge at the point of diffraction (see Figure 1). The half-angle of the cone is just the angle between the incident ray and the edge. When the incident ray is perpendicular to the edge, the half-angle is $\pi/2$ and the cone of diffracted rays becomes a plane perpendicular to the edge.

The 'law of edge diffraction' and the straightness of the diffracted rays are consequences of an extended form of Fermat's principle[1]. The special form of this principle which suffices in the present problem is the following: A singly diffracted ray connecting two points P and Q is a curve whose length is stationary among all curves connecting these two points and having one point on the edge of the screen. From this principle it follows that a diffracted ray consists of two straight line segments, one from P and one from Q to some point on the edge. Furthermore these lines must make equal angles with the edge and lie on opposite sides of the plane normal to the edge at the point of diffraction. This result is the law of edge diffraction stated above.

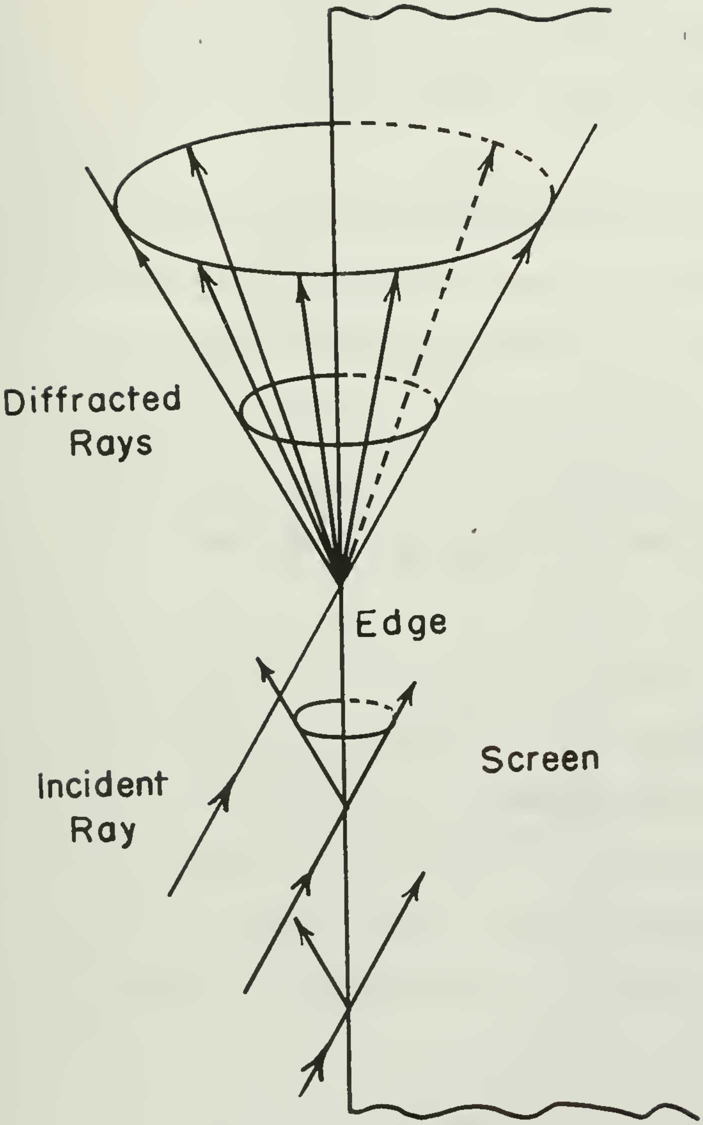


Figure 1a.

The cone of diffracted rays produced by an incident ray which hits the edge of a thin screen.

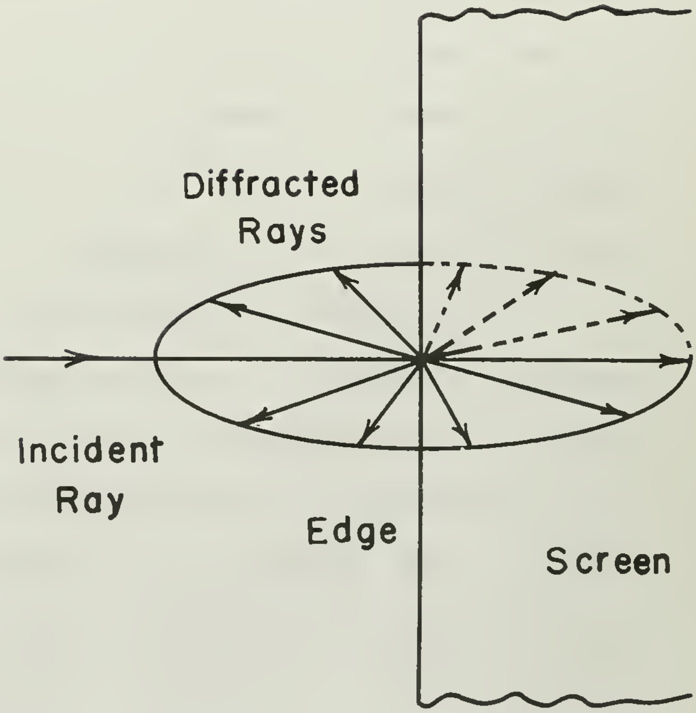


Figure 1b.

The plane of diffracted rays produced by a ray normally incident on the edge of a thin screen.

We also assume that a diffracted ray which hits the edge can again produce diffracted rays in accordance with the'law of edge diffraction'. Thus doubly, triply and multiply diffracted rays can occur. All of these rays also follow from the extended Fermat's principle in which curves with more than one point on the edge are considered.

As an example of the preceding considerations Figure 2 shows the rays produced by a plane wave normally incident upon a thin screen containing a slit. Since each incident ray which hits an edge of the screen is perpendicular to this edge, the resulting diffracted rays lie in a plane normal to the edge. Through each point behind the screen there pass two singly diffracted rays (one from each edge), two doubly diffracted rays, etc. In addition an incident ray passes through each point not in the shadow.

So far we have not considered any ray which hits a corner or vertex of the edge. We assume that such rays also give rise to diffracted rays which are straight lines satisfying the 'law of vertex diffraction.' This law states that a ray diffracted from a vertex may have any direction, without any relation to the direction of the incident ray. As a consequence of this law it follows that a single ray which hits a vertex produces diffracted rays which leave the vertex in all directions.

The straightness of the vertex diffracted rays and the law of vertex diffraction also follow from the extended form of Fermat's principle[1]. The special form of this principle which characterizes the vertex diffracted rays is this: A ray connecting two points P and Q and singly diffracted from the vertex V is a curve whose length is stationary among all curves connecting these two points and passing through V. This principle implies that such a ray consists of two straight line segments, one from P to V and one from V to Q. The directions of these two lines are unrelated.

A ray diffracted from a vertex may hit another vertex or an edge and

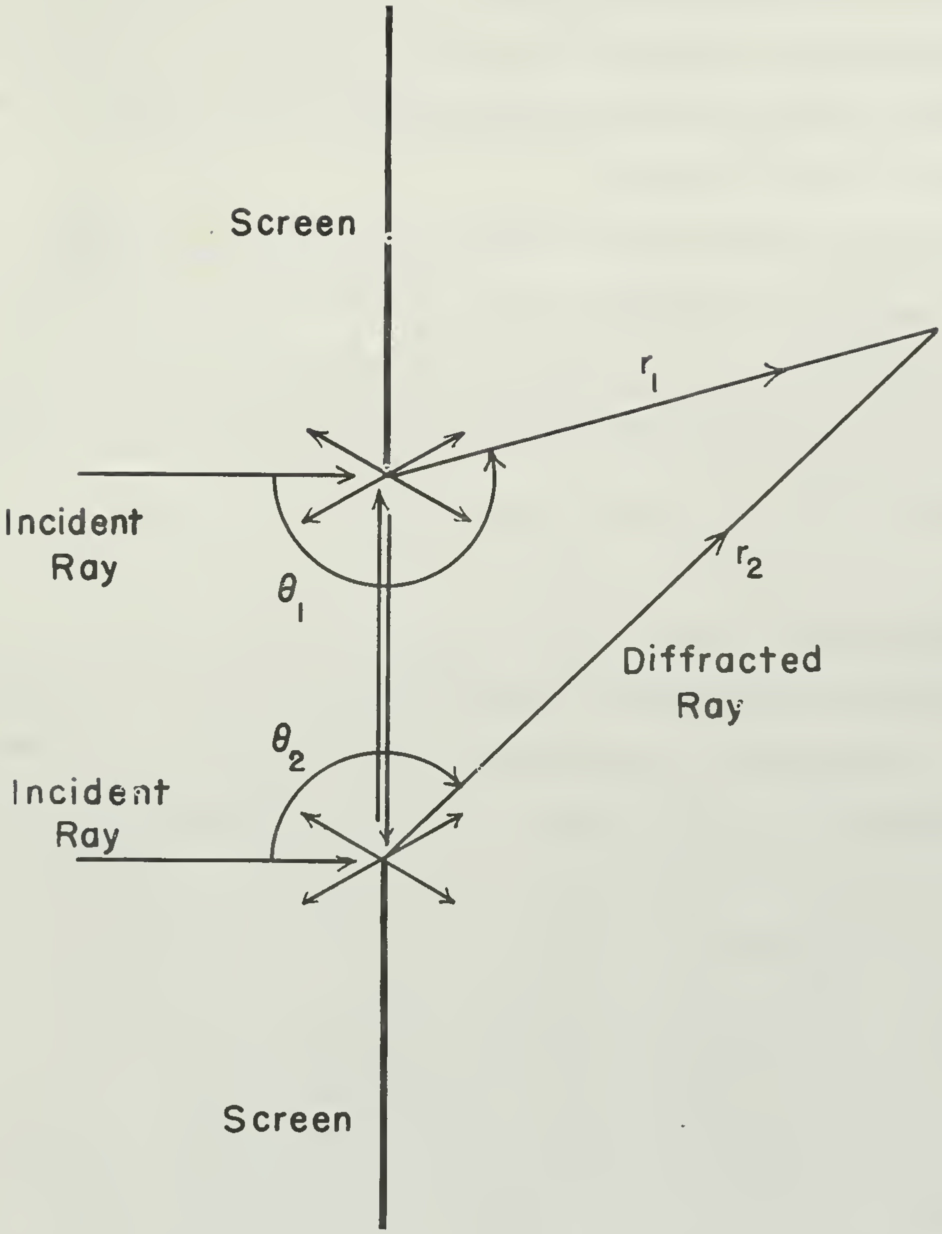


Figure 2

The diffracted rays produced by a plane wave normally incident upon a slit in a thin screen. The two incident rays which hit the slit edges are shown along with some of the singly diffracted rays which they produce. One diffracted ray from each edge is shown crossing the slit and hitting the opposite edge, producing doubly diffracted rays.

produce other diffracted rays. These rays may in turn produce other diffracted rays, etc. All of the diffracted rays--both singly and multiply diffracted from edges and vertices--must be taken into account. They can all be characterized by a more general extension of Fermat's principle which is similar to the two special forms previously described.

One or more diffracted rays pass through each point in space. These rays account for the field in the shadow, and also account for modifications of the field outside the shadow. As we shall see, the singly diffracted rays are quantitatively most important and the edge diffracted rays more so than the vertex diffracted ones.

3. The field on a ray

Now that the rays have been characterized, we turn to the quantitative description of the field. To this end we assume that there is an amplitude A and a phase $k\psi$ associated with each ray. Both of these quantities may vary with the distance s (from some origin) along the ray. In terms of these quantities the field on the ray is assumed to be given by $A(s)e^{ik\psi(s)}$. If the field is electromagnetic, A is a vector, but if the field is scalar--which we will now suppose--then A is a scalar. In neither case is A assumed to be real.

We assume that the difference in phase between two points on the same ray is equal to k times the length of the portion of the ray between them. Thus for two points at s and s_0 we have $\psi(s) - \psi(s_0) = s - s_0$. If we choose $s_0 = 0$ and denote $\psi(0)$ by ψ_0 this yields

$$(2) \qquad \psi(s) = \psi_0 + s.$$

We also assume that there is a direction of propagation on each ray and that s increases in this direction. The direction of propagation thus determines the direction in which the phase increases. From these considerations we see that the phase can be found at any point s on a ray by means of equation (2) provided that it is known at some other point on the same ray and provided that the direction of propagation is also known.

To determine how the amplitude varies along a ray, we assume that the optical form of the principle of conservation of energy is satisfied. This principle states that the flux of energy is the same at every cross section of a tube of rays in a non-dissipative medium, such as the one under consideration. Furthermore we assume that the flux through a narrow tube of cross sectional area $d\sigma$ is $cA^2 d\sigma$ where c is the propagation velocity. By applying this energy principle to two cross sections of such a tube, we obtain the relation $cA^2 d\sigma = c_o A_o^2 d\sigma_o$. In a homogeneous medium the propagation velocity is constant, so $c = c_o$. Thus we find from the energy relation that $A = A_o \left(\frac{d\sigma_o}{d\sigma} \right)^{1/2}$

If we imagine that the tube shrinks down upon a single ray then this equation relates the amplitudes A and A_o at two points -- say P and P_o -- on this ray. In this relation $d\sigma_o/d\sigma$ is the limit of the area ratio as the tube shrinks to the ray. This limit is just the Jacobian of the mapping of one cross section of the tube on the other by means of rays. Thus we find that the amplitude A can be computed at any point P on a ray provided the amplitude A_o is known at some other point P_o on the same ray. To perform this computation it is not sufficient to deal with one ray, but it is instead necessary to take account of other rays near the one under consideration in order to determine the factor $d\sigma_o/d\sigma$.

The area ratio $d\sigma_o/d\sigma$ can be computed very easily if we first observe that a normal cross section of a narrow tube of rays is just a portion of a wavefront. Then the two cross sections at the two points along the tube are portions of parallel wavefronts. Now the ratio of areas of corresponding small portions of parallel surfaces is the reciprocal of the ratio of the Gaussian curvatures of these surfaces. To compute this ratio we denote by ρ_1 and ρ_2 the principal radii of curvature of the wavefront at the point P_0 on the ray. Then the corresponding radii at the point P a distance s along the ray are ρ_1+s and ρ_2+s. Thus, since the Gaussian curvature is the reciprocal of the product of the principal radii, we have*

$$(3) \qquad \frac{d\sigma_0}{d\sigma} = \frac{\rho_1\rho_2}{(\rho_1+s)(\rho_2+s)} .$$

Now using this in our preceding expression for A we obtain

$$(4) \qquad A = A_0 \left(\frac{\rho_1\rho_2}{\left[\rho_1+s\right]\left[\rho_2+s\right]} \right)^{1/2} .$$

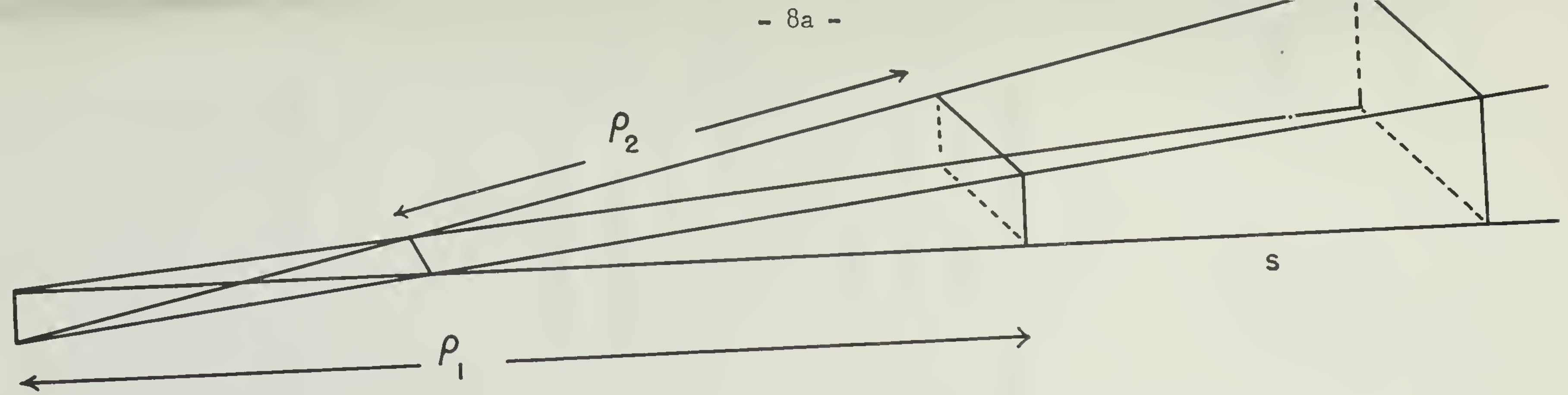
In order to combine the expressions (2) and (4), which determine the phase and amplitude on a ray, we first choose P_0 as the origin of s. Then at the point s on the ray the field u is given by

$$(5) \qquad u = A_0 \left[\frac{\rho_1\rho_2}{(\rho_1+s)(\rho_2+s)} \right]^{1/2} e^{ik(s+\psi_0)} .$$

From (5) we see that for large s, the field on a ray decreases like s^{-1}, as in a spherical wave, provided that both ρ_1 and ρ_2 are finite. If one radius is infinite, the field decays like $s^{-1/2}$, as in a cylindrical wave, while if both

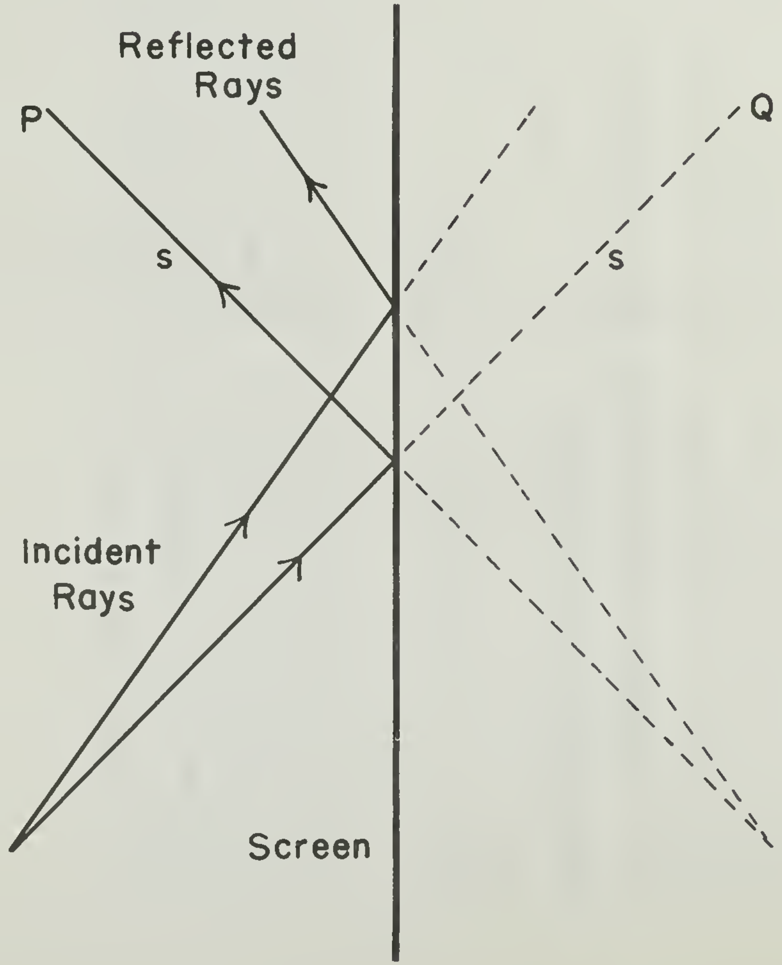
---

[*] This expression for $d\sigma_0/d\sigma$ follows at once from the geometrical interpretation of the radii of curvature shown in Figure 3. Note that the ratio of corresponding horizontal lines in the two wavefronts is $\rho_1/(\rho_1+s)$ and that of corresponding vertical lines is $\rho_2/(\rho_2+s)$. The ratio of the areas is the product of these two ratios.



↑ Figure 3

Two small portions of parallel wavefronts a distance s apart and some of the rays normal to them. The two centers of curvature of these wavefronts, where the neighboring rays intersect, are at the distances ρ_1 and ρ_2 from one wavefront, and at the distances ρ_1+s and ρ_2+s from the other.



<- Figure 4

A pair of neighboring incident rays hitting a plane screen and the corresponding reflected rays. The distance from a point P on a reflected ray to the center of curvature is the same as the distance from its image point Q, on the extension of the incident ray, to the corresponding center of curvature.

radii are infinite the field is of constant amplitude, as in a plane wave. In these respective cases there are two, one, or no points on the ray at which the above expression for the field becomes infinite. Such points are called caustic points and their locus is the caustic associated with the ray system. Since the present theory yields an infinite value for the field on a caustic, it is obviously not valid at such a surface. Later we shall see how this defect of the theory can be remedied.

4. Reflection and diffraction coefficients

The preceding considerations have resulted in equation (5), which enables us to find the field at any point on a ray provided that it is known at one point on the ray. We will now determine the field at the point of initiation of each reflected and diffracted ray, assuming that it is known on the corresponding incident ray. To this end we assume that the initial value of the field on each reflected or diffracted ray is proportional to the value of the field on the corresponding incident ray at the point of reflection or diffraction. The respective coefficients of proportionality will be called reflection and diffraction coefficients.

Applying these considerations to the field u_{ref} on a reflected ray, we have at the point of reflection

$$(6) \qquad u_{ref}(0,y,z) = R\, u_{inc}(0,y,z).$$

In (6) R is the reflection coefficient. This coefficient depends upon the nature of the field and the nature of the screen at the point of reflection. It may also depend upon the angle of incidence and upon k. As an example, if the total field $u_{inc} + u_{ref}$ vanishes on the screen, then $R = -1$, as we see from (6). On the other hand, if the normal derivative of the total field vanishes, then we find that $R = +1$. Other boundary conditions lead to other values of R.

Equation (6) determines the initial value of the field on each reflected ray. Therefore, with the aid of equation (5), the field can be found at all points on each reflected ray. To calculate this field it is necessary to determine the radii of curvature of the reflected wavefronts at each point of reflection on the screen. Because the screen is plane, these radii are equal to the corresponding radii of the incident wavefronts at each point. Therefore the geometrical factor $\left[\rho_1\rho_2/(\rho_1+s)(\rho_2+s)\right]^{1/2}$ and the phase factor $e^{ik(s+\Psi_0)}$ will have the same values at a point P distant s from the screen along a reflected ray as they would have at a point Q the same distance behind the screen along the corresponding incident ray (see Figure 4). But these factors would give the value of the incident field at Q. Therefore, since the y- and z-coordinates of P and Q are the same, while their x-coordinates differ in sign, (5) and (6) yield

$$(7) \qquad u_{ref}(x,y,z) = R\,u_{inc}(-x,y,z).$$

In (7) R is the value of the reflection coefficient appropriate to the reflected ray through x,y,z. If no reflected ray passes through x,y,z then $u_{ref}(x,y,z) = 0$.

Let us now consider the field on an edge-diffracted ray. First we note that neighboring diffracted rays of the same cone intersect each other at the diffracting edge. Thus the edge is a caustic of the set of diffracted rays and each point of it is a center of curvature of the diffracted wavefronts. Therefore, as we have seen, our result (5) will yield an infinite value for the field at the edge. Consequently we cannot determine the field on a diffracted ray simply by equating the diffracted field at the edge to a multiple of the incident field. Instead we proceed as follows.

Consider the point P_o from which the distance s is measured along a diffracted ray. Let ρ_2 be the distance from P_o to the edge along the ray. Then ρ_2 is one radius of curvature of the diffracted wavefront at P_o, because one center of curvature of the wavefront is at the edge. Now rewrite (4) in the form

$$(8) \qquad A_o\, \rho_2^{1/2} = A\left[(\rho_1+s)(\rho_2+s)\right]^{1/2} \cdot \rho_1^{-1/2}$$

If the right-hand side of (8) is evaluated at some fixed point P on the diffracted ray, then it is independent of the location of P_o. This is so because A is a constant at P and because ρ_1+s and ρ_2+s are the distances from P along the ray to the two centers of curvature of the diffracted wavefront. Therefore the left side of (8) is also a constant independent of the location of P_o. Let us denote this constant by A_o'. Furthermore, let us choose P_o at the edge. Then $\rho_2 = 0$ and (8) yields

$$(9) \qquad A = A_o'\left[\frac{\rho_1}{(\rho_1+s)s}\right]^{1/2}.$$

In (9), s denotes distance from the edge along a diffracted ray.

By combining (9) and (2) we obtain for the field $u_{\text{edge diff}}$ on a ray diffracted from an edge, the result

$$(10) \qquad u_{\text{edge diff}} = A_o'\left[\frac{\rho_1}{(\rho_1+s)s}\right]^{1/2} e^{ik(s+\psi_o)}.$$

The radius of curvature ρ_1 is the distance from the edge to the other caustic of the diffracted rays. This distance is computed in Appendix I, and is found to be given by (see Figure 5)

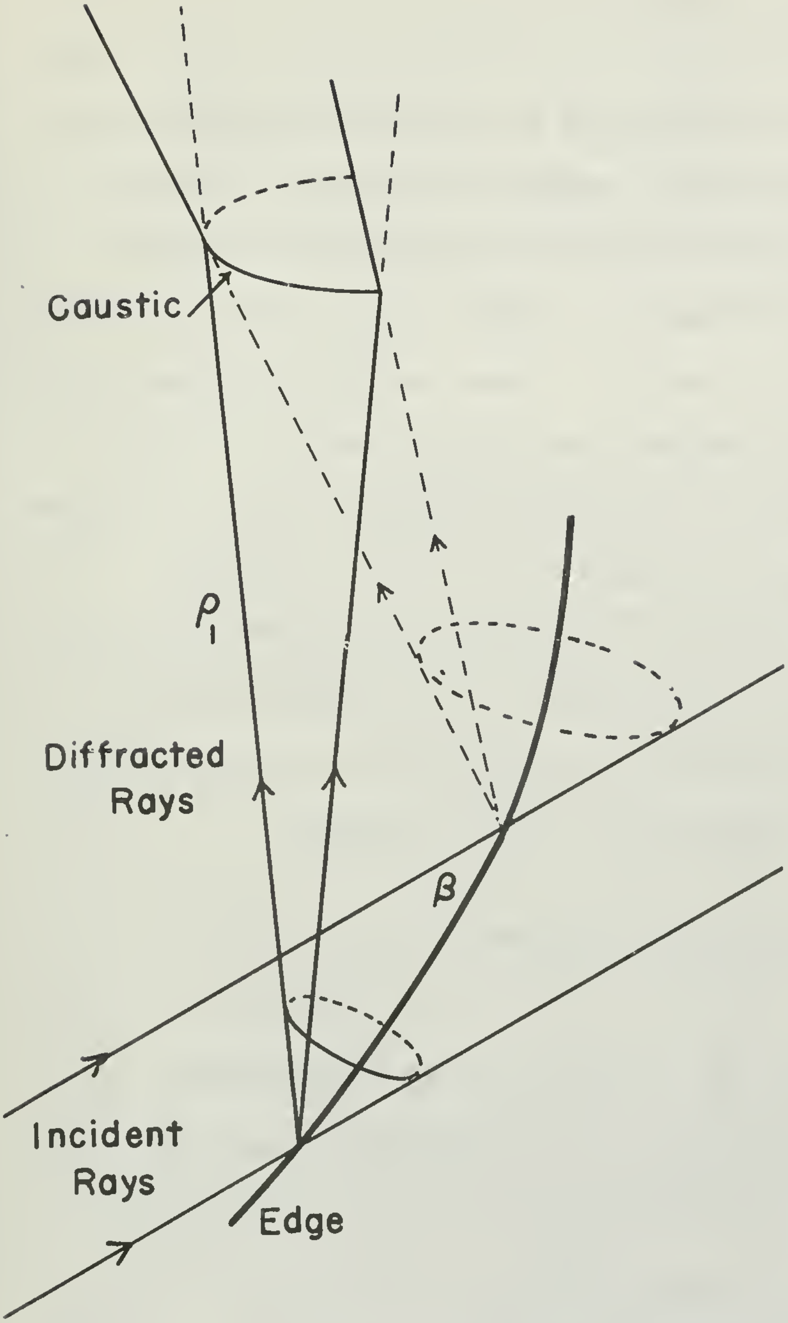


Figure 5

A pair of neighboring incident rays hitting the curved edge of a thin screen, and some of the diffracted rays. The two cones of diffracted rays intersect at the caustic, which is at the distance ρ_1 from the edge along one of the diffracted rays.

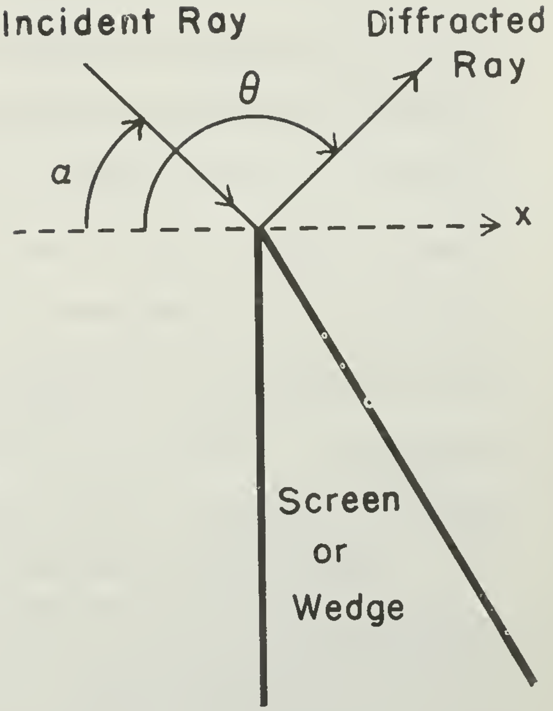


Figure 6

The projection of incident and diffracted rays into a plane normal to the edge. The angles α and θ are the angles between these projections and the negative x-axis, measured as shown in the figure. The diffracting edge, which is perpendicular to the plane of the figure, is the edge of the screen or the edge of a wedge (see Appendix II).

$$(11) \qquad \rho_1 = -\rho \sin^2\beta \, (\rho \dot{\beta} \sin\beta + \cos\delta)^{-1}.$$

In (11), $\rho \geq 0$ denotes the radius of curvature of the edge, β is the angle between the incident ray and the (positive) tangent to the edge, $\dot{\beta}$ is the derivative of β with respect to arclength along the edge, and δ is the angle between the diffracted ray and the normal to the edge. It is to be noted that equation (11) is independent of the direction in which arclength increases along the edge since both factors $\dot{\beta}$ and $\sin\beta$ change sign when this direction is reversed. Furthermore the normal to the edge, which lies in the plane of the screen, is supposed to point toward the center of curvature of the edge.

We now assume that the factor $A_o' e^{ik\psi_o}$, which determines the field on a diffracted ray according to (10), is proportional to the incident field $Ae^{ik\psi}$ at the point of diffraction. We will denote the coefficient of proportionality by D and call it the diffraction coefficient. Thus (10) becomes

$$(12) \qquad u_{\text{edge diff}} = D \, A \left[s(1+\rho_1^{-1} s) \right]^{-1/2} e^{ik(s+\psi)} \quad .$$

The diffraction coefficient depends upon the nature of the field and the nature of the screen at the point of diffraction. It also may depend upon k and upon the directions of the incident and diffracted rays. It is to be noted that $u_{\text{edge diff}}$ behaves like a cylindrical wave ($s^{-1/2}$) near the edge, and like a spherical wave (s^{-1}) far from the edge, provided that ρ_1 is finite.

In order to obtain two examples of diffraction coefficients, we will now consider a scalar field which satisfies the reduced wave equation. In the first case we suppose that the field vanishes on the screen and in the second case that its normal derivative vanishes on the screen. In each case the diffraction of a plane wave by a screen in the form of a half-plane can be described exactly by the solution of a boundary-value problem for the reduced wave equation. This was shown by Sommerfeld[5]. These exact solutions can also be expanded asymptotically for large k. On the other hand, these problems can be treated by the present method, using (7) for the reflected field and (12) for the diffracted field. The asymptotic expansion of the exact solution and the field computed from (7) and (12) are compared in Appendix II and are found to agree perfectly provided that

$$D = -\frac{e^{i\pi/4}}{2\sqrt{2\pi k}\,\sin\beta}\left[\sec\tfrac{1}{2}(\theta-\alpha) \pm \csc\tfrac{1}{2}(\theta+\alpha)\right]. \tag{13}$$

The angles α and θ in (13) are defined as follows: Project the incident and diffracted rays into the plane normal to the edge at the point of diffraction. The angles between these projections and the normal to the screen are α and θ respectively (see Figure 6.) The angle β is the angle between the incident ray and the edge. The upper sign in (13) applies to the case in which the field vanishes on the screen and the lower sign to the case in which the normal derivative vanishes on the screen.

Equation (13) gives the diffraction coefficients for two kinds of screens for a scalar field satisfying the reduced wave equation. Thus for such fields and screens the field on an edge diffracted ray is, from (11), (12) and (13),

$$u_{\text{edge diff}} = -\frac{Ae^{ik(\psi+s)+i\pi/4}}{2\sqrt{2\pi k}\,\sin\beta}\left[\sec\frac{1}{2}(\theta-\alpha) \pm \csc\frac{1}{2}(\theta+\alpha)\right] \cdot \left[s\left(1 - \frac{s[\cos\delta + \rho\dot{\beta}\sin\beta]}{\rho\sin^2\beta}\right)\right]^{-1/2}. \tag{14}$$

It is to be noted that D, and therefore the edge-diffracted field given by (14), are inversely proportional to $\sqrt{k}$. Thus the diffracted field decreases in magnitude as k increases, i.e. as λ decreases. This is in accordance with our expectations about diffraction. This behavior is quite general, and is not restricted to the special fields or screens which we have just considered. It follows from (12) that D must have the dimension $(\text{length})^{1/2}$ since A and $u_{\text{edge diff}}$ have the same dimensions. The only length upon which D may depend is λ and therefore D must be proportional to $\lambda^{1/2}$ or to $k^{-1/2}$.

Let us turn now to the corner-diffracted rays. Since these rays emanate from the corner in all directions, the corresponding diffracted wavefronts are spheres with the corner as center. Both centers of curvature of these wavefronts are at the corner, which is consequently a caustic. Considerations similar to those which led from (5) to (12) in the case of edge diffraction now lead from (5) to

$$(15) \qquad u_{\text{corner diff}} = C\,A\,\frac{e^{ik(s+\Psi)}}{s} \quad .$$

Here s denotes distance from the corner, $Ae^{ik\Psi}$ is the incident field at the corner, and C is the coefficient of corner diffraction. C, like D and R, depends upon the nature of the field, the nature of the screen at the corner, the directions of the incident and diffracted rays, and the value of k. In addition it depends upon the angle between the two edges meeting at the corner. From (15) we see by dimensional considerations that C must be proportional to λ or k^{-1}. Thus the corner-diffracted field decreases faster than the edge-diffracted field as k increases.

In order to determine C for a scalar field satisfying the reduced wave equation and for screens on which the field or its normal derivative vanishes, we may proceed as follows. Consider a plane wave incident upon an infinite corner, i.e. a screen with two semi-infinite straight edges meeting at a corner. Solve this boundary-value problem exactly for the reduced wave equation and expand the solution asymptotically for large k. Compare this result with the field given by (7), (14) and (15). The two results should agree for an appropriate choice of C, which will then be the desired corner-diffraction coefficient. This procedure has not yet been carried out because of the intractable form of the solution of the corner-diffraction problem, which has been considered by L. Kraus[6].

5. The total field

We have now determined all the rays, found how the field varies along each ray and determined the 'initial' value of the field on each reflected or singly diffracted ray. Exactly the same formulas apply to a doubly diffracted ray, the corresponding singly diffracted ray being considered as the incident ray. In the same way the field on every multiply diffracted ray can be determined. We are now ready to consider the total field.

We assume that the total field u at any point P is the sum of the fields on all the rays through P. We also define the diffracted field at P, u_{diff}, as the sum of the fields on all diffracted rays through P. Finally we define the geometrical optics field at P, $u_{geom} = u_{inc} + u_{ref}$ as the sum of the fields on all the incident and reflected rays through P. Then we can write

$$u = u_{geom} + u_{diff} \tag{16}$$

The geometrical optics field is zero in the shadow region, through which neither incident nor reflected rays pass. Thus in the shadow the total field is u_{diff}. However u_{diff} is also present in the 'illuminated' region, in which u_{geom} is not zero.

The diffracted field u_{diff} consists of a finite or infinite sum of the fields on all the singly, doubly and multiply diffracted rays through each point. These rays may be diffracted from edges, from corners, or from both. Each time a ray is diffracted by an edge, its field is multiplied by a diffraction coefficient D which contains the factor $k^{-1/2}$. Each time it is diffracted by a corner, its field is multiplied by a corner diffraction coefficient C which contains the factor k^{-1}. Thus for large values of k, the fields on multiply diffracted rays become rapidly weaker as the number of diffractions increases. For very large values of k, the singly edge-diffracted field, $u_{\text{single edge diff}}$, is most significant, all the other diffracted fields being much weaker. Even the singly corner-diffracted field is insignificant compared to it. Of course, in those regions of the shadow, if any, where $u_{\text{single edge diff}}$ is zero, $u_{\text{single corner diff}}$ becomes important.

The singly edge-diffracted field at any point P can be written as

$$(17) \qquad u_{\text{single edge diff}}(P) = \sum D\, A \left[s(1+\rho_1^{-1} s)\right]^{-1/2} e^{ik(s+\Psi)}.$$

The sum extends over all the singly edge-diffracted rays through P. In case D is given by (13), (17) becomes

$$(18) \qquad u_{\text{single edge diff}}(P) = -\sum \frac{Ae^{ik(\Psi+s)+i\pi/4}}{2\sqrt{2\pi k}\,\sin\beta} \left[\sec\frac{1}{2}(\theta-\alpha) \pm \csc\frac{1}{2}(\theta+\alpha)\right] \cdot \left[s\left(1 - \frac{s[\cos\delta + \rho\dot{\beta}\sin\beta]}{\rho\sin^2\beta}\right)\right]^{-1/2}.$$

The formula (18) expresses the singly diffracted field explicitly in terms of geometrical quantities. This formula will be applied in the following sections to apertures and disks of various special shapes. In certain cases the corresponding formula which includes all multiply diffracted rays will also be used.

Before we apply our result to special cases, we will make certain general comments concerning it. In the first place, it is clear from the derivation that the result (12) applies to diffraction by any edge, and not merely by the edge of a thin screen. For example, it applies to diffraction by the edge of a wedge, provided that the appropriate diffraction coefficient is used. This coefficient is deduced for two kinds of wedges -- those on which the field vanishes and those on which its normal derivative vanishes -- in Appendix II. Similarly, equation (15) applies to diffraction by any corner or vertex, and not merely to a corner on the edge of a thin screen. Thus it applies to diffraction by the tip of a cone or the corner of a cube, provided that the appropriate diffraction coefficient is used in each case.

Next it is important to observe that the diffraction coefficients can always be obtained from certain exact special or canonical solutions of boundary-value problems for the differential equations governing the field. The geometry in the canonical problem may be very simple, provided only that it coincides locally with the geometry near the point of diffraction. Therefore it may involve just a plane incident wave and a wedge-like or conical obstacle.

Canonical solutions can also be used to correct the results of the present theory at and near caustics and shadow boundaries, where the present theory yields infinite values for the field. Thus, for example, at an edge the diffracted field is infinite. Furthermore the geometrical optics field is discontinuous across the shadow boundary which passes through the edge. Near the edge these two fields may be replaced by the field obtained from the solution of a boundary-value problem for diffraction by a half-plane. The edge of the half-plane must be tangent to the actual aperture edge. Of course the incident field and the properties of the screen in the half-plane problem must correspond to those in the actual problem. This procedure is just the one which was proposed by Braunbek[7] as a basis for the calculation of diffraction by an aperture in a thin screen.

In Part II[8] of this article we will show that his proposal can be worked out in detail for an aperture of any shape and that it leads exactly to our equation (18). It then follows that our method is equivalent to Braunbek's as far as rays singly diffracted from edges are concerned. However his methods does not take account of multiple diffraction, nor does it yield the correct field on corner-diffracted rays. Furthermore, it requires an integration in order to compute the field, while the present method does not.

In Part II we will also consider the various forms of the Kirchhoff method and show that each of them leads to a result of the form given by equation (17). However the diffraction coefficients D will not coincide with the exact diffraction coefficients given by equation (13).

It should be observed that the diffraction coefficients given by equation (13) become infinite on the shadow boundary $\theta = -\alpha$ and on the boundary of the reflected field $\theta = \pi + \alpha$. These are the surfaces across which the geometrical optics field $u_{inc} + u_{ref}$ is discontinuous. Thus the present results are invalid on these surfaces.

The modifications of the preceding results which must be made in the case of electromagnetic fields are rather minor. If the field u denotes an electric or a magnetic field then it must be a vector. We assume that all of our preceding results apply to these fields when we consider the amplitude A to be a vector whose direction remains constant along a ray and which is perpendicular to it. If we consider both electric and magnetic fields, they must also be perpendicular to each other. Since A is a vector, the reflection coefficient R and the diffraction coefficients D and C must now be matrices. The matrices R for reflection of the electric or magnetic field from a perfect conductor or from a dielectric are known. These matrices are given by the well-known exact solutions of Maxwell's equations for the problems of reflection of a plane electromagnetic wave from a plane conductor and from a plane dielectric interface. Consequently they will not be given here. However in Appendix III we consider the diffraction of an electromagnetic wave by the edge of a perfectly conducting screen and determine the diffraction matrix D.

We will now apply our results to various problems, at first assuming the field to be scalar and the diffraction coefficient D to be given by (13).

6. Diffraction of spherical and cylindrical waves by a half-plane

As a first application of our result, let us find the field produced when an arbitrary wave is incident upon a half-plane. In this case no multiply diffracted or corner-diffracted rays occur. We will assume that the angle between the incident ray and the edge varies monotonically along the edge, as is the case with incident plane, spherical and cylindrical waves. Then there is just one diffracted ray through each point not on the edge. Thus u_{diff} consists of a single term which is given by equation (12) or (14). These formulas can be simplified since ρ, the radius of curvature of the edge, is infinite. It then follows from (11) that $\rho_1 = -\dot{\beta}^{-1} \sin\beta$.

If the incident wave is cylindrical and if its axis is parallel to the edge, then $\beta = \pi/2$ for all rays incident on the edge. The field in a cylindrical wave is of the form $a_o \frac{e^{ikr}}{\sqrt{kr}}$, where r denotes distance from the axis and a_o is a constant. This follows from arguments similar to those used in deriving equation (12). We choose $a_0 = \sqrt{\frac{2}{\pi}}\, e^{-i\pi/4}$, so that the cylindrical wave is that due to a line source of unit strength. Now if r_0 is the distance of the line source from the edge, then upon inserting the above expression for the cylindrical wave into (14) we obtain

$$(19) \qquad u_{diff} = -\frac{e^{ik(s+r_0)}}{2\pi k\sqrt{r_0 s}}\left[\sec\frac{1}{2}(\theta-\alpha) \pm \csc\frac{1}{2}(\theta+\alpha)\right].$$

Equation (19) gives the field due to diffraction of a cylindrical wave from a screen on which u vanishes, if the upper sign is taken, or on which the normal derivative of u vanishes, if the lower sign is taken. The boundary-value problem for the reduced wave equation can be solved exactly for such screens with a cylindrical incident wave, and the solutions can be expanded asymptotically for large k. The results of these expansions are precisely given by equation (19) plus u_{geom}. This agreement constitutes a check on our theory.

Next let us suppose that the incident wave is spherical, with its center at a point O. Then if r denotes distance from O, the incident field is of the form $a_0 e^{ikr}/r$ and we will set $a_0 = 1$. This form of the field can be obtained by arguments like those which lead to (15). At a point Q on the edge at distance r from O, $\dot{\beta}^{-1} = -r/\sin\beta$ where β is the angle between the ray from O to Q. Since $\rho_1 = -\dot{\beta}^{-1}\sin\beta$, we see that $\rho_1 = r$ at Q. Using these results in (14) we obtain

$$(20) \qquad u_{diff} = -\frac{e^{ik(r+s)+i\pi/4}}{2\sin\beta\sqrt{2\pi k}\sqrt{rs(r+s)}}\left[\sec\frac{1}{2}(\theta-\alpha) \pm \csc\frac{1}{2}(\theta+\alpha)\right].$$

Equation (20) gives the field due to diffraction of a spherical wave from a screen on which u or its normal derivative vanishes, according as the upper or lower sign is taken. The boundary-value problem for the reduced wave equation has been solved exactly for such screens with a spherical wave incident. The solution can be expanded asymptotically for large k and the result agrees precisely with u_{diff} in equation (20) plus u_{geom}. This agreement is another check on our theory.

7. Diffraction by a slit

As a second application of our result we will compute the field produced by a plane wave incident upon a slit of width 2a and of infinite length, in a thin screen (see Figure 2). We will suppose that each incident ray lies in a plane normal to the edges of the slit, so that $\beta = \pi/2$ for all rays incident upon the edges. This problem is essentially two-dimensional, since each diffracted ray will also lie in a plane normal to the edges. The more general case in which $\beta \neq \pi/2$ can be treated almost as easily, or the results for it can be obtained from the present ones by the method explained in Appendix II.

As we have seen, two diffracted rays of each order will pass through each point. To compute the field on the diffracted rays we make use of the facts that the radius of curvature of each edge is infinite and that $\beta = \pi/2$, so $\dot{\beta} = 0$. It then follows that ρ_1 is also infinite. Now using (18) we can write the singly diffracted field provided that we know the incident field. Let us assume that the incident field is $e^{ik(x \cos \alpha - y \sin \alpha)}$. Then (18) yields

$$(21) \qquad u_{\text{single diff}} = - \frac{e^{ik(r_1 - a \sin \alpha) + i\pi/4}}{2\sqrt{2\pi k r_1}} \left[\sec \tfrac{1}{2}(\theta_1 + \alpha) \pm \csc \tfrac{1}{2}(\theta_1 - \alpha) \right]$$

$$- \frac{e^{ik(r_2 + a \sin \alpha) + i\pi/4}}{2\sqrt{2\pi k r_2}} \left[\sec \tfrac{1}{2}(\theta_2 - \alpha) \pm \csc \tfrac{1}{2}(\theta_2 + \alpha) \right].$$

In equation (21) r_1 and r_2 denote the distances of a point from the upper edge at $(0, +a)$ and from the lower edge at $(0, -a)$ respectively. The angles θ_1 and θ_2 are determined by the corresponding rays, as shown in Figure 2.

Before computing the fields on the multiply diffracted rays, we will examine the singly diffracted field far from the slit. To this end we consider the polar coordinates r, ϕ of a point in the x,y-plane. If this point is far from the slit then we see that $r_1 \sim r - a \sin \phi$, $r_2 \sim r + a \sin \phi$, $\theta_1 \sim \pi + \phi$, and $\theta_2 \sim \pi - \phi$. Upon using these expressions in equation (21) we obtain

$$(22) \qquad u_{\text{single diff}} = -\sqrt{\frac{k}{2\pi r}}\, e^{ikr+i\pi/4} f_s(\phi)$$

In equation (22) we have introduced $f_s(\phi)$, the radiation or diffraction pattern of the slit due to singly diffracted rays. From (21) it is found to be given by

$$(23) \qquad f_s(\phi) = i\,\frac{\sin\left[ka\,(\sin\phi + \sin\alpha)\right]}{k \sin\frac{1}{2}\,(\phi + \alpha)} \pm \frac{\cos\left[ka\,(\sin\phi + \sin\alpha)\right]}{k \cos\frac{1}{2}\,(\phi - \alpha)} .$$

In Figures 7 and 8, $|k\, f_s(\phi)|$ is shown as a function of ϕ for normal incidence ($\alpha = 0$). The angle ϕ ranges between 0 and $\pi/2$, corresponding to the region behind the slit. Figure 7 is based upon the value $ka = 8$, corresponding to a slit about 3 wavelengths wide, and Figure 8 corresponds to a slit 10 wavelengths wide, for which $ka = 10\pi$. It is to be noted that the number of maxima in each curve equals the number of wavelengths which fit into the slit. Furthermore it is interesting that the curves are the same for both the upper and lower signs in (23), i.e. for both the screen on which $u = 0$ and that on which the normal derivative of u is zero.

For comparison the curve based upon the exact solution of the boundary-

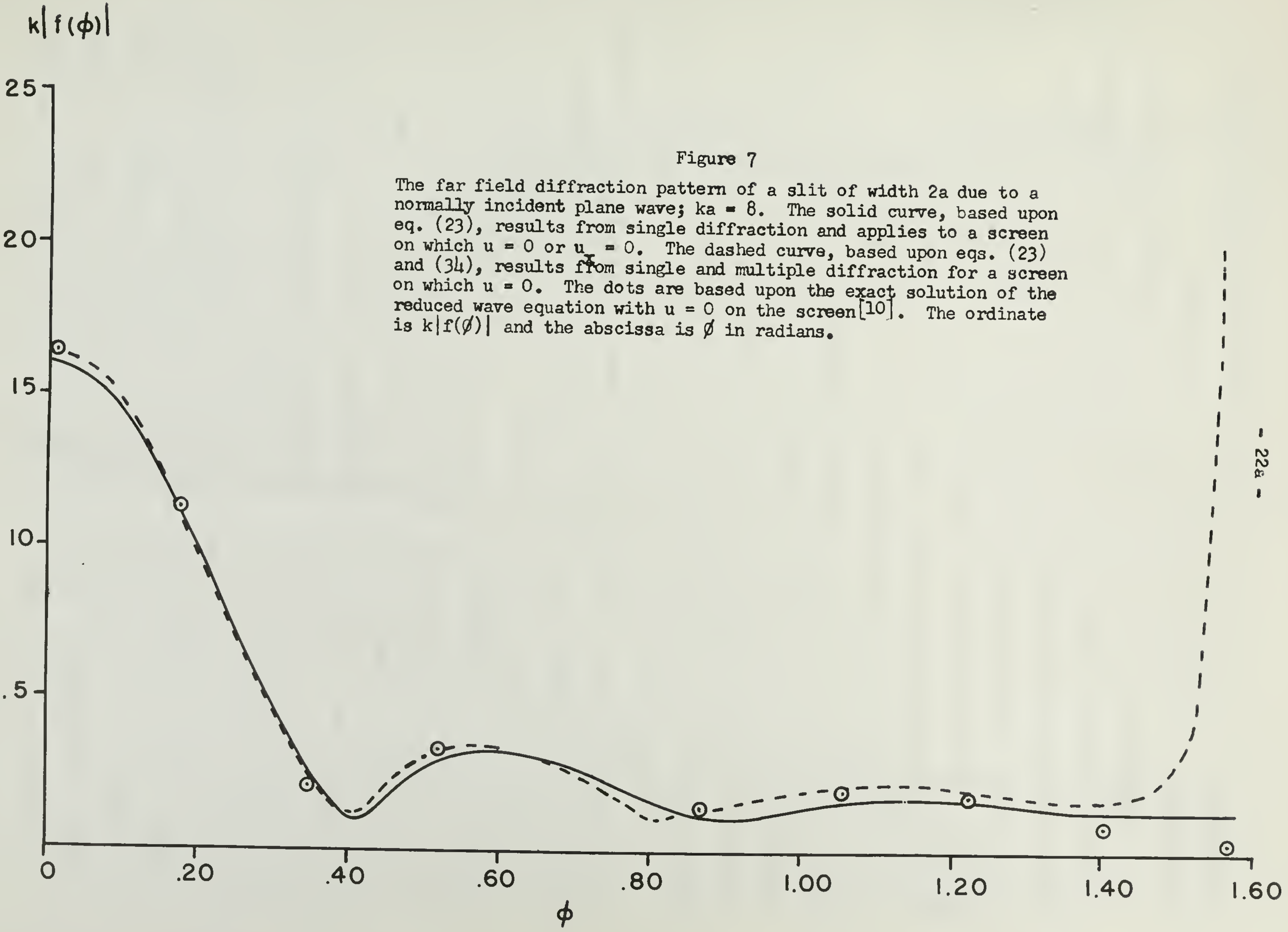


Figure 7

The far field diffraction pattern of a slit of width 2a due to a normally incident plane wave; ka = 8. The solid curve, based upon eq. (23), results from single diffraction and applies to a screen on which u = 0 or $u_x = 0$. The dashed curve, based upon eqs. (23) and (34), results from single and multiple diffraction for a screen on which u = 0. The dots are based upon the exact solution of the reduced wave equation with u = 0 on the screen[10]. The ordinate is $k|f(\phi)|$ and the abscissa is ϕ in radians.

value problem for the reduced wave equation is also shown in Figure 7. This curve applies to the boundary condition $u = 0$ on the screen.

According to the cross section theorem, the transmission cross section of the slit (per unit length) is equal to the imaginary part of the diffraction pattern in the forward direction $\phi = -\alpha$. Applying this theorem to $f_s(\phi)$ we find that the transmission cross section is $2a \cos \alpha$. Since this is the projection of the slit width on the direction normal to the incident rays, it is just the geometrical optics cross section. Thus the singly diffracted rays do not contribute to the cross section when it is computed from $f_s(\phi)$ by the cross section theorem. Of course they would contribute if the cross section were computed by integration of $f_s(\phi)$ with respect to ϕ. This discrepancy does not contradict the cross section theorem because the theorem applies only to the exact solution of the boundary-value problem for the reduced wave equation.

The above calculation of the cross section did not take explicit account of u_{geom} nor of the fact that $u_{single\ diff}$ becomes infinite on the two boundaries of the shadow. Actually u_{geom}, which is discontinuous on the shadow boundaries, should merge with $u_{single\ diff}$ as r increases. This merging is automatically accounted for in the calculation of $u_{single\ diff}$ for r large. In this calculation the singularities on the two shadow boundaries cancel each other and the resulting finite expression includes the effect of u_{geom}. This is shown by the cross section result. However in other problems the same simple calculations do not suffice and other considerations are needed. This will be seen in the case of the circular aperture.

In order to obtain a first correction to the geometrical optics cross section we need merely take account of the two doubly diffracted rays which go in the direction $\phi = -\alpha$. To do this we note that the singly diffracted ray which leaves the lower edge in the direction $\theta_2 = \pi/2$ hits the upper edge. It reaches the upper edge when $r_2 = 2a$. Thus the field on this ray at the upper edge is,

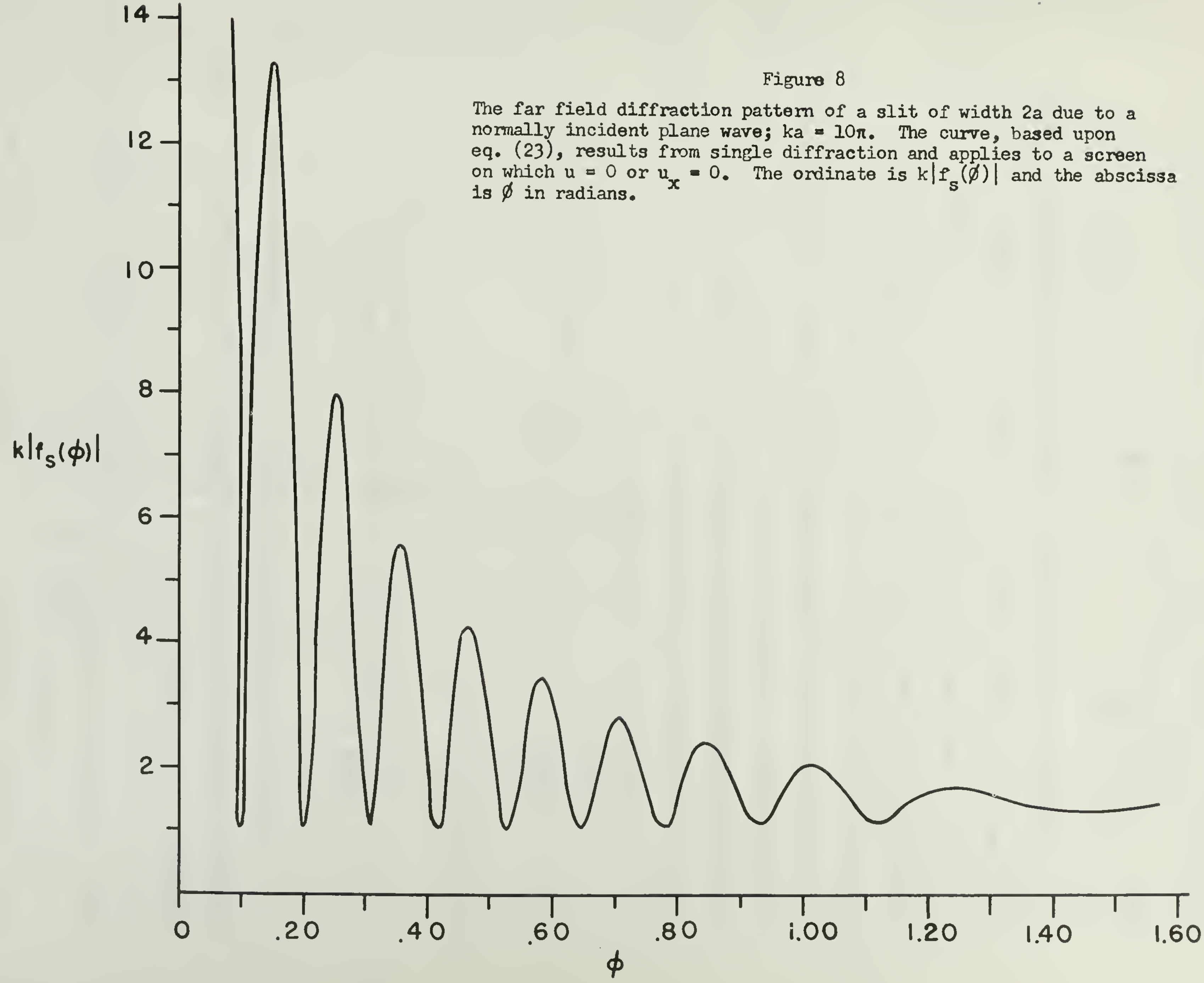


Figure 8

The far field diffraction pattern of a slit of width 2a due to a normally incident plane wave; $ka = 10\pi$. The curve, based upon eq. (23), results from single diffraction and applies to a screen on which $u = 0$ or $u_x = 0$. The ordinate is $k|f_s(\phi)|$ and the abscissa is ϕ in radians.

if we choose the upper sign in (21),

$$(24) \qquad -\frac{e^{ik(2a+\sin\alpha)+i\pi/4}}{2\sqrt{\pi ka}} \sec\frac{1}{2}\left(\frac{\pi}{2}-\alpha\right)$$

If we had chosen the lower sign in (21) rather than the upper one, we would have found that the field on the ray $\theta_2 = \pi/2$ is zero. Therefore the field on all doubly and multiply diffracted rays would be zero in this case, and (22) would give the complete diffracted field. We will reconsider this case later, but now we will proceed with our calculation for the screen on which $u = 0$.

To find the field on the doubly diffracted rays produced by diffraction at the upper edge, we again use equation (14), treating (24) as the incident field. We proceed in a similar way to find the field on the doubly diffracted rays produced at the lower edge. Combining the fields on the two doubly diffracted rays in the direction $\phi = -\alpha$ we obtain

$$(25) \qquad \underset{\text{single diff}}{u(r,-\alpha)} + \underset{\text{double diff}}{u(r,-\alpha)} = -\sqrt{\frac{k}{2\pi r}}\, e^{ikr+i\pi/4}\left[f_s(-\alpha) + f_d(-\alpha)\right].$$

Here $f_d(-\alpha)$ is defined by

$$(26) \qquad f_d(-\alpha) = -\frac{1}{k\sqrt{\pi ka}}\left[\frac{e^{i\left[2ka(1+\sin\alpha)+\pi/4\right]}}{1+\sin\alpha} + \frac{e^{i\left[2ka(1-\sin\alpha)+\pi/4\right]}}{1-\sin\alpha}\right].$$

We now apply the cross section theorem, which requires that we take the imaginary part of $f_s(-\alpha) + f_d(-\alpha)$. We thus obtain for the transmission cross section σ the expression

$$(27) \qquad \sigma = 2a\cos\alpha - \frac{1}{k\sqrt{\pi ka}}\left[\frac{\cos\left[2ka(1+\sin\alpha)-\pi/4\right]}{1+\sin\alpha} + \frac{\cos\left[2ka(1-\sin\alpha)-\pi/4\right]}{1-\sin\alpha}\right].$$

The correction to the geometrical optics cross section is of the order $k^{-3/2}$. This result applies to a slit in a screen on which u = 0. The result (27) has also been obtained by H. Levine[9]. In the special case of normal incidence, $\alpha = 0$ and (27) becomes

$$\frac{\sigma}{2a} = 1 - \frac{\cos(2ka - \pi/4)}{\sqrt{\pi}\,(ka)^{3/2}}. \tag{28}$$

A graph of $\sigma/2a$ as a function of ka is shown in Figure 9, based upon equation (28). For comparison a curve based upon the exact solution of the boundary-value problem for the reduced wave equation is also shown.

We will now determine the field on each multiply diffracted ray. Then we will add all these fields together to obtain the total multiply diffracted field. For the reason previously stated, only the case in which u = 0 on the screen will be considered.

The different kinds of multiply diffracted rays are classified in the following table:

Type	Initial Diffraction	Number of diffractions at each edge Lower	Upper	Final Diffraction
A	Lower	n + 1	n	Lower
B	Lower	n	n	Upper
C	Upper	n	n + 1	Upper
D	Upper	n	n	Lower

The first column designates the type of ray. The second column states which edge is first hit by the incident ray that gives rise to a particular multiply diffracted ray. The last column states the edge at which the multiply diffracted ray is finally diffracted. The other columns give the numbers of times the ray is diffracted at each edge. The integer n is at least unity, since the rays are all multiply diffracted.

To determine the field on any ray we make repeated use of equation (14). The calculation is simplified if we introduce the function $F(s,\alpha,\theta)$ defined by

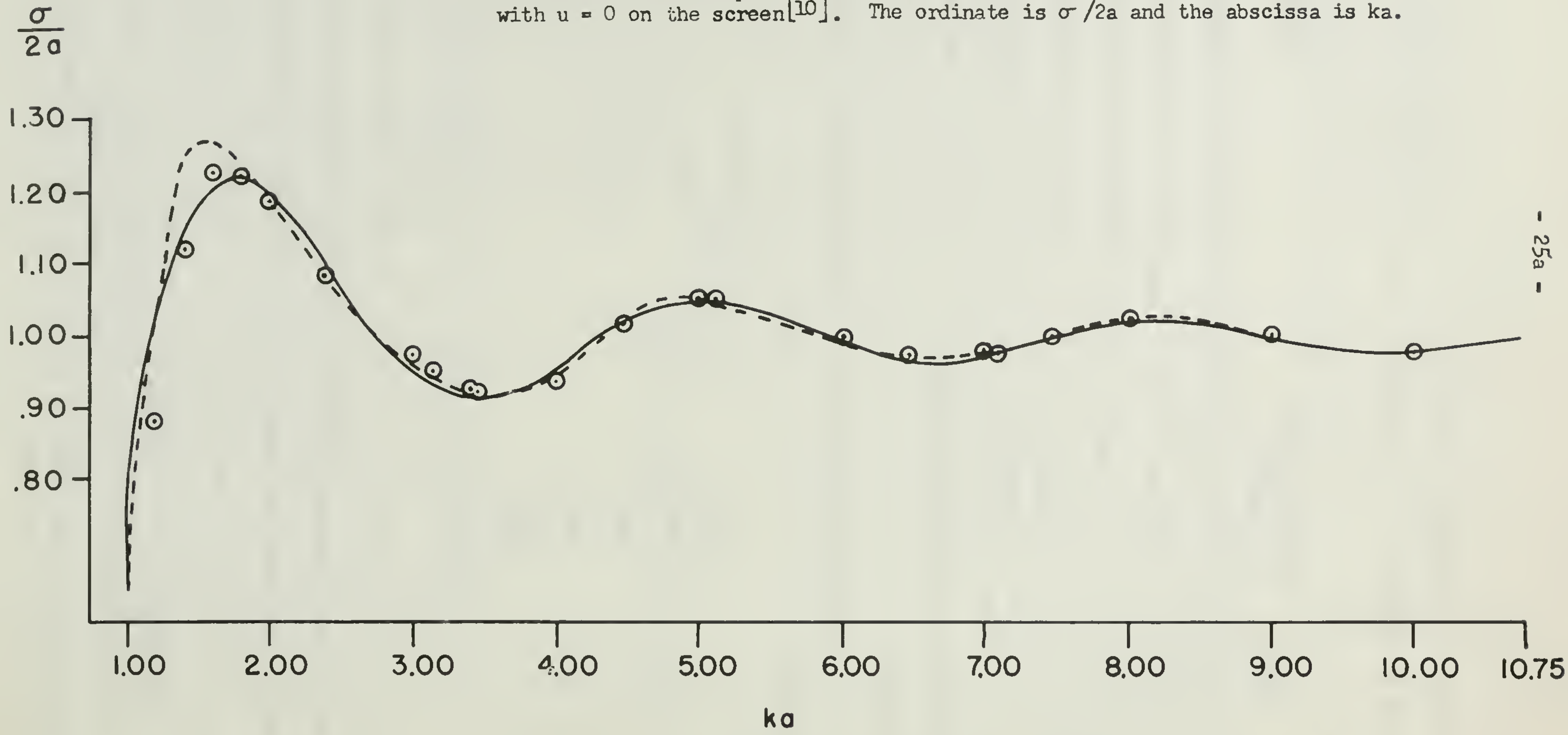


Figure 9

The transmission cross section of a slit of width 2a as a function of ka, for normal incidence with u = 0 on the screen. The solid curve, based upon eq. (28), results from single and double diffraction; the dashed curve based upon eq. (35) results from single and all multiple diffraction. The dots are based upon the exact solution of the reduced wave equation with u = 0 on the screen[10]. The ordinate is $\sigma/2a$ and the abscissa is ka.

$$(29) \qquad F(s,\alpha,\theta) = -\frac{e^{iks+i\pi/4}}{2\sqrt{2\pi ks}}\left[\sec\frac{1}{2}(\theta-\alpha)+\csc\frac{1}{2}(\theta+\alpha)\right].$$

Then according to equation (14), the diffracted field in the present problem is simply the incident field at the edge multiplied by $F(s,\alpha,\theta)$. Thus, for example, the field (24) at the upper edge on the ray singly diffracted at the lower edge is just $e^{ika\sin\alpha}F(2a,\alpha,\frac{\pi}{2})$. By referring to the table above, one easily finds that the fields on the various rays are given by

$$(30) \qquad \begin{aligned} &A)\ e^{ika\sin\alpha}F(2a,\alpha,\tfrac{\pi}{2})\,F^{2n-1}(2a,\tfrac{\pi}{2},\tfrac{\pi}{2})\,F(r_2,\tfrac{\pi}{2},\theta_2),\\ &B)\ e^{ika\sin\alpha}F(2a,\alpha,\tfrac{\pi}{2})\,F^{2n-2}(2a,\tfrac{\pi}{2},\tfrac{\pi}{2})\,F(r_1,\tfrac{\pi}{2},\theta_1),\\ &C)\ e^{-ika\sin\alpha}F(2a,-\alpha,\tfrac{\pi}{2})\,F^{2n-1}(2a,\tfrac{\pi}{2},\tfrac{\pi}{2})F(r_1,\tfrac{\pi}{2},\theta_1),\\ &D)\ e^{-ika\sin\alpha}F(2a,-\alpha,\tfrac{\pi}{2})\,F^{2n-2}(2a,\tfrac{\pi}{2},\tfrac{\pi}{2})F(r_2,\tfrac{\pi}{2},\theta_2). \end{aligned}$$

The total multiply diffracted field $u_{\text{mult diff}}$ is now obtained by adding together the fields on the four types of rays and summing with respect to n. The only series to be summed is the geometric series $\sum_{n=1}^{\infty} F^{2n-2}(2a,\pi/2,\pi/2)$ which has the sum $\left\{1-F^2(2a,\pi/2,\pi/2)\right\}^{-1}$. Thus we find

$$(31) \qquad u_{\text{mult diff}} = \left\{ e^{ika\sin\alpha}F(2a,\alpha,\pi/2)\left[F(2a,\pi/2,\pi/2)\,F(r_2,\pi/2,\theta_2)+F(r_1,\pi/2,\theta_1)\right]\right.$$
$$\left. + e^{-ika\sin\alpha}F(2a,-\alpha,\pi/2)\left[F(2a,\pi/2,\pi/2)F(r_1,\pi/2,\theta_1)+F(r_2,\pi/2,\theta_2)\right]\right\}$$
$$\left\{1-F^2(2a,\pi/2,\pi/2)\right\}^{-1}.$$

The complete diffracted field $u_{\text{diff}} = u_{\text{single diff}} + u_{\text{mult diff}}$ is given by equations (21) and (31). Exactly the same result has been obtained by S.N. Karp and A. Russek[10] by a very different method.

The expression (31) for the multiply diffracted field can be considerably simplified at points far from the slit. By using the notation and approximations explained just preceding equation (22), we find that (31) simplifies to

$$(32) \qquad u_{\text{mult diff}} = -\sqrt{\frac{k}{2\pi r}}\, e^{ikr+i\pi/4}\, f_m(\phi).$$

The radiation or diffraction pattern $f_m(\phi)$ due to multiple diffraction is given by

$$(33) \qquad f_m(\phi) = \left[4\pi ikae^{-i4ka}+1\right]^{-1} \Bigg\{ -\frac{1}{k}\Big[\sec\left(\frac{\pi}{4}-\frac{\alpha}{2}\right)\sec\left(\frac{\pi}{4}-\frac{\phi}{2}\right) e^{ika(\sin\alpha+\sin\phi)}$$

$$+\sec\left(\frac{\pi}{4}+\frac{\alpha}{2}\right)\sec\left(\frac{\pi}{4}+\frac{\phi}{2}\right)e^{-ika(\sin\alpha+\sin\phi)}\Big]$$

$$+2\sqrt{\frac{\pi a}{k}}\, e^{-i2ka-i\frac{\pi}{4}}\Big[\sec\left(\frac{\pi}{4}-\frac{\alpha}{2}\right)\sec\left(\frac{\pi}{4}+\frac{\phi}{2}\right)e^{ika(\sin\alpha-\sin\phi)}$$

$$+\sec\left(\frac{\pi}{4}+\frac{\alpha}{2}\right)\sec\left(\frac{\pi}{4}-\frac{\phi}{2}\right)e^{-ika(\sin\alpha-\sin\phi)}\Big]\Bigg\}.$$

when $\alpha = 0$ equation (33) simplifies to

$$(34) \qquad f_m(\phi) = -\frac{4}{k}\left[\sqrt{4\pi ka}\, e^{-i2ka-i\frac{\pi}{4}}+1\right]^{-1} \frac{\cos\frac{\phi}{2}\cos(ka\sin\phi)-i\sin\frac{\phi}{2}\sin(ka\sin\phi)}{\cos\phi}.$$

A graph of k times the total diffraction pattern, $\left|k\left\{f_s(\phi)+f_m(\phi)\right\}\right|$, is shown in Figure 7 for normal incidence ($\alpha = 0$) with $ka = 3\pi$.

To find the transmission cross section of the slit, we again apply the cross section theorem. Now we must compute the imaginary part of $f_s(-\alpha) + f_m(-\alpha)$.

In the case of normal incidence ($\alpha = 0$) this yields

$$(35) \qquad \frac{\sigma}{2a} = 1 - \frac{\cos\left(2ka - \frac{\pi}{4}\right)}{\sqrt{\pi}\,(ka)^{3/2}} \left[1 - (\pi ka)^{-1/2} \sin\left(2ka - \frac{\pi}{4}\right) + (4\pi ka)^{-1}\right]^{-1}$$

A graph of $\sigma/2a$ as a function of ka for normal incidence ($\alpha=0$) is shown in Figure 9, based upon equation (35). Corresponding curves based upon equation (28) and upon an exact solution of the reduced wave equation are also shown. It can be seen that both the result (28) based upon double diffraction and equation (35) are quite close to the exact curve. However, the curve based upon (35) is closer, especially for ka small, where multiple diffraction is important.

From (33) we see that $f_m(\phi)$ contains as a factor the reciprocal of $(4\pi i ka\, e^{-i4ka} + 1)$. If this quantity is zero, then $f_m(\phi)$ and the field u_m will not be defined. The values of k for which this occurs are all complex, and correspond to the free vibrations or modes of the field. These free vibrations or modes contain no incident waves, but consist entirely of outgoing waves. Thus they must decay exponentially in time, and we see that this is so from their angular frequencies $\omega = kc$. The corresponding factor in (34) is the reciprocal of $(\sqrt{4\pi ka}\, e^{-i2ka-i\frac{\pi}{4}} + 1)$. This quantity is a factor of the previous one, and thus contains only some of the roots. These roots must correspond to modes symmetric about the x-axis since they occur in the case of normal incidence. The other modes are consequently antisymmetric. All the roots of this equation have negative imaginary parts, corresponding to decaying modes.

8. Diffraction by a circular aperture

We will now apply our results to diffraction of a normally incident plane wave by a circular aperture. In this case two singly diffracted rays pass through each point which is neither on the edge nor on the axis of the aperture. They come to any point P from the nearest and farthest points on the edge. Thus the singly diffracted field given by (18) contains just two terms. To evaluate them we note that $\beta = \pi/2$ for all incident rays, so $\dot{\beta} = 0$. If the radius of the hole is a, then $\rho = a$. Assuming that the incident wave is e^{ikx} we obtain from (18)

$$u_{\text{single diff}}(P) = -\frac{e^{ikr_1+i\pi/4}}{2\sqrt{2\pi k}}\left[\sec\frac{\theta_1}{2} \pm \csc\frac{\theta_1}{2}\right]\left[r_1(1-a^{-1}r_1\sin\theta_1)\right]^{-1/2}$$

(36)

$$-\frac{e^{ikr_2+i\pi/4}}{2\sqrt{2\pi k}}\left[\sec\frac{\theta_2}{2} \pm \csc\frac{\theta_2}{2}\right]\left[r_2(1-a^{-1}r_2\sin\theta_2)\right]^{-1/2}$$

In (36) r_1 and r_2 denote the distances from P to the two points of diffraction on the edge, while θ_1 and θ_2 are the angles determined by the corresponding rays. In writing (36) we used the relations $\delta_1 = \theta_1 - \frac{\pi}{2}$ and $\delta_2 = \theta_2 - \frac{\pi}{2}$.

To simplify (36) we consider a point P far from the hole and introduce the polar coordinates r, ϕ in the plane containing P and the axis. In this plane the ray configuration and the angles are the same as in the case of normal incidence on a slit, which is shown in Figure 2. Therefore the approximations made in that case, preceding equation (22), are applicable again and (36) becomes

$$(37) \qquad u_{\text{single diff}} = -\frac{ke^{ikr}}{2\pi r} f_s(\phi).$$

The radiation or diffraction pattern $f_s(\phi)$ due to single diffraction is given by

$$(38) \qquad f_s(\phi) = k^{-3/2}\sqrt{\frac{2\pi a}{\sin\phi}}\left[i\,\frac{\sin(ka\sin\phi - \frac{\pi}{4})}{\sin\frac{\phi}{2}} \pm \frac{\cos(ka\sin\phi - \frac{\pi}{4})}{\cos\frac{\phi}{2}} \right]$$

A graph of $|kf_s(\phi)|$ as a function of ϕ, based upon (38), is shown in Figure 10 for $ka = 3\pi$. As we see from (38) and from the graph, $f_s(\phi)$ becomes infinite at $\phi = 0$. This is because the axis $\phi = 0$ is a caustic of the singly diffracted rays and because the shadow boundary $r = a$ also extends in the direction $\phi = 0$.

Let us now examine the field (36) at points near the axis. We set $\rho = (y^2+z^2)^{1/2}$ and consider points for which ρ is small. For such points we set $\theta_1 = \frac{\pi}{2} + \delta_1$ and $\theta_2 = \frac{\pi}{2} + \delta_2$ where $\delta_1 = \tan^{-1}x/(a-\rho)$ and $\delta_2 = \tan^{-1}x/(a+\rho)$. Then $r_1 = x\sin\delta_1 + (a-\rho)\cos\delta_1$ and $r_2 = x\sin\delta_2 + (a+\rho)\cos\delta_2$. Thus $1 - a^{-1}r_1\sin\theta_1 = a^{-1}\rho$ and $1 - a^{-1}r_2\sin\theta_2 = -a^{-1}\rho$. Finally we note that $r_1 \approx r_2 \approx (x^2+a^2)^{1/2}$ and that $\delta_1 \approx \delta_2 \approx \delta = \tan^{-1}x/a$. Inserting these results into (36) yields

$$(39) \qquad u_{\text{single diff}} = -\frac{\sqrt{a}\, e^{ik(x\sin\delta + a\cos\delta)}}{\sqrt{2\pi k\rho}\,(x^2+a^2)^{1/4}}\left[\sec\left(\frac{\delta}{2} + \frac{\pi}{4}\right) \pm \csc\left(\frac{\delta}{2} + \frac{\pi}{4}\right)\right]$$

$$\cdot \cos\left(k\rho\cos\delta - \frac{\pi}{4}\right).$$

The expression (39) for $u_{\text{single diff}}$ becomes infinite on the axis $\rho = 0$. This is due to the fact that the axis is a caustic of the singly diffracted rays. In order to correct (39) so that it yields a finite value on the axis, we make use of an exact solution of the reduced wave equation which has an axial caustic. This is done in Appendix IV ; it turns out that (39) can be corrected on and near the axis by multiplying it by $\frac{1}{2}\sqrt{2\pi k\rho\cos\delta}\,\sec(k\rho\cos\delta - \frac{\pi}{4})\,J_0(k\rho\cos\delta)$. J_0 is the Bessel function of zero order. Then (39) becomes

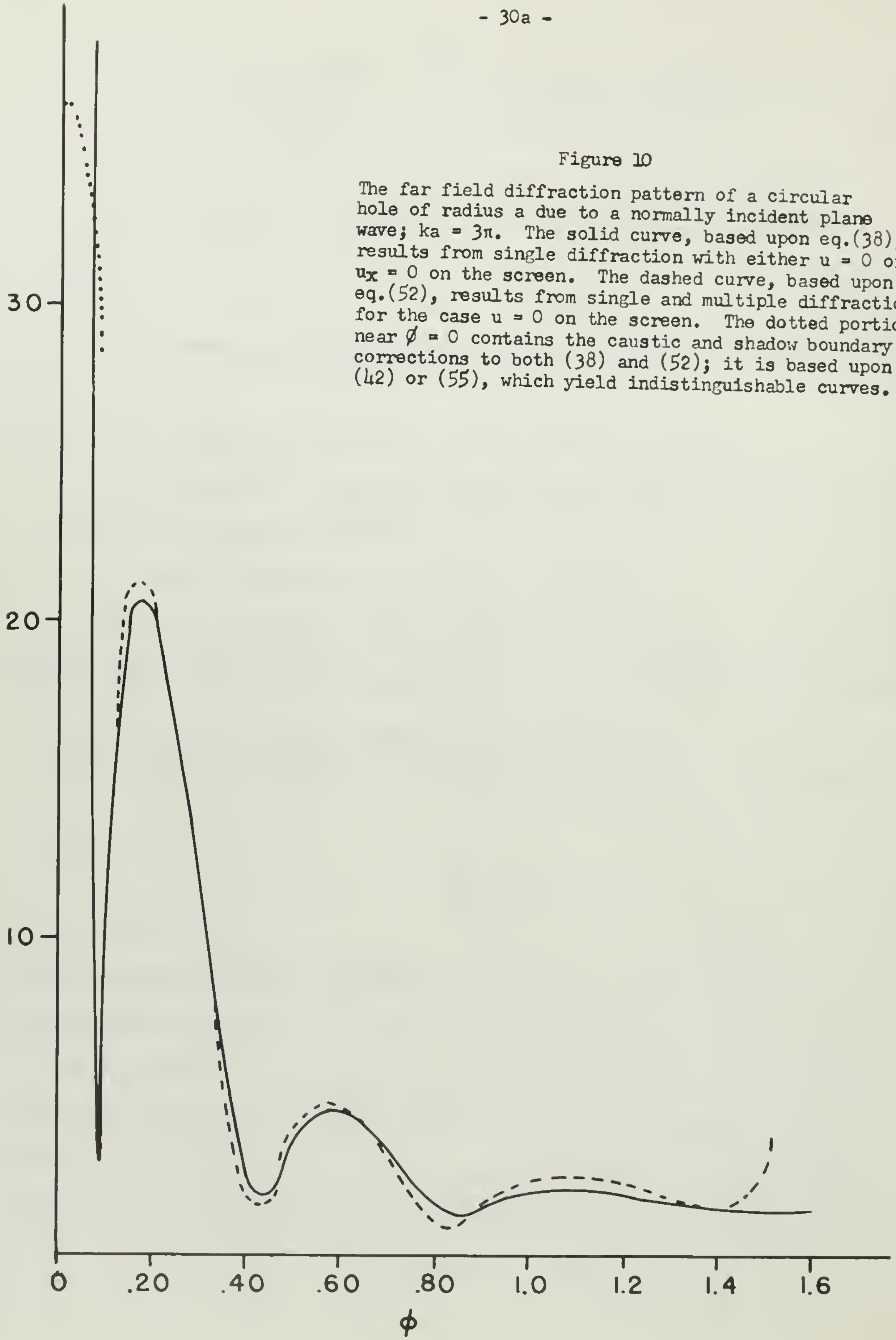


Figure 10

The far field diffraction pattern of a circular hole of radius a due to a normally incident plane wave; ka = 3π. The solid curve, based upon eq.(38), results from single diffraction with either u = 0 or u_x = 0 on the screen. The dashed curve, based upon eq.(52), results from single and multiple diffraction for the case u = 0 on the screen. The dotted portion near ϕ = 0 contains the caustic and shadow boundary corrections to both (38) and (52); it is based upon (42) or (55), which yield indistinguishable curves.

$$(40)\qquad u_{\text{single diff}} = -\frac{\sqrt{a\cos\delta}}{2(x^2+a^2)^{1/4}}\left[\sec\left(\frac{\delta}{2}+\frac{\pi}{4}\right) \pm \csc\left(\frac{\delta}{2}+\frac{\pi}{4}\right)\right] J_o(k\rho\cos\delta)e^{ik(x^2+a^2)^{1/2}}$$

which is finite on and near the axis.

Far from the screen, $\cos\delta \approx a/x$ and then (40) becomes (with $\sin\phi = \rho/x$)

$$(41)\qquad u_{\text{single diff}} = -\frac{a}{2x}\exp\left[ik\left(x+\frac{a^2}{2x}\right)\right]\left[\sec\left(\frac{\delta}{2}+\frac{\pi}{4}\right) \pm \csc\left(\frac{\delta}{2}+\frac{\pi}{4}\right)\right] J_o(ka\sin\phi).$$

We cannot use (41) directly to compute the transmission cross section of the aperture because δ approaches $\pi/2$ as x becomes large, and then $\sec\left(\frac{\delta}{2}+\frac{\pi}{4}\right) \approx 2x/a$ becomes infinite. This is due to the fact that the shadow boundary extends in the direction $\delta = \frac{\pi}{2}$, which is parallel to the axis. Instead we first combine $u_{\text{single diff}}$ with u_{geom}, which is e^{ikx} in the lit region. If we write $\exp\left[ik\left(x+\frac{a^2}{2x}\right)\right] \approx e^{ikx}\left(1+\frac{ika^2}{2x}\right)$ then we find from (41) that

$$(42)\qquad u_{\text{geom}} + u_{\text{single diff}} \approx e^{ikx} - \frac{a}{2x}e^{ikx}\left(1+\frac{ika^2}{2x}\right)\frac{2x}{a}J_o(ka\sin\phi)$$

$$= -\frac{ke^{ikx}}{2\pi x}\cdot i\pi a^2 J_o(ka\sin\phi).$$

From (42) we see that u_{geom} and $u_{\text{single diff}}$ combine near the axis and far from the screen to give a spherical wave with a finite coefficient. Thus $f_s(\phi)$ given by (38) is to be replaced by $i\pi a^2 J_o(ka\sin\phi)$ for ϕ near zero. According to the cross section theorem, the transmission cross section is the imaginary part of $f_s(0)$, which is now just πa^2. This is the geometrical optics cross section.

In order to obtain a first correction to the geometrical optics cross section, we will take account of doubly diffracted rays. Each incident ray gives rise to a diffracted ray in the plane of the screen. This ray is normal to the edge, and therefore it goes through the center of the aperture and is normally incident on the opposite edge. The field on this ray, when it reaches the opposite edge, is given by, say, the first term in (36) with $r_1 = 2a$, and $\theta_1 = \frac{\pi}{2}$. This field is

$$(43) \qquad - \frac{e^{2ika - i\pi/4}}{4\sqrt{\pi ka}} (\sqrt{2} \pm \sqrt{2}) .$$

From (43) we see that this field is zero for a screen on which the normal derivative of the field vanishes as was pointed out previously. Thus we will consider for the present only screens on which the field itself vanishes.

If we treat (43) as the incident field, (18) yields

$$(44) \qquad u_{\text{double diff}} = \frac{e^{ik(r_1 + 2a)}}{2\pi k\sqrt{a}} \sec\frac{\delta_1}{2} \left[r_1(1 - a^{-1}r_1\cos\delta_1) \right]^{-1/2}$$

$$+ \frac{e^{ik(r_2 + 2a)}}{2\pi k\sqrt{a}} \sec\frac{\delta_2}{2} \left[r_2(1 - a^{-1}r_2\cos\delta_2) \right]^{-1/2} .$$

From (44) and (39) we see that near the axis

$$(45) \qquad u_{\text{double diff}} = - \frac{2e^{2ika - i\frac{\pi}{4}}}{\sqrt{2\pi ka}} \sec\frac{\delta}{2} \left[\sec\left(\frac{\delta}{2} + \frac{\pi}{4}\right) + \csc\left(\frac{\delta}{2} + \frac{\pi}{4}\right) \right]^{-1} u_{\text{single diff}} .$$

Therefore $u_{\text{double diff}}$ becomes infinite on the axis just like $u_{\text{single diff}}$ does, because the axis is also a caustic of the doubly diffracted rays. It can be corrected near the axis in the same way as before. The result is easily obtained by using (41) in (45), which yields for points near the axis and far from the screen,

$$(46)\qquad u_{\text{double diff}} = -\frac{ke^{ikx}}{2\pi x}\left(-\frac{2}{k}\sqrt{\frac{\pi a}{k}}\right) e^{2ika-i\pi/4} J_0(ka \sin \phi).$$

Applying the cross section theorem to (46), and adding in the cross section πa^2 obtained from (42) we have

$$(47)\qquad \frac{\sigma}{\pi a^2} = 1 - \frac{2 \sin(2ka - \frac{\pi}{4})}{\sqrt{\pi}\,(ka)^{3/2}} .$$

The correction to the geometrical cross section is of order $(ka)^{-3/2}$ and is oscillatory. The result (47) has also been obtained by H. Levine[11].

A graph of $\sigma/\pi a^2$ as a function of ka is shown in Figure 11, based upon (47). For comparison values of $\sigma/\pi a^2$ computed from an exact solution of the boundary-value problem for the reduced wave equation are also shown.

We will now calculate the multiply diffracted field $u_{\text{mult diff}}$. This calculation is almost exactly the same as that for the slit, since each multiply diffracted ray is normal to the edge which it hits. In fact the previous calculation applies identically if we replace the function $F(s,\alpha,\theta)$ in (29) by

$$(48)\qquad F(s,\alpha,\theta) = -\frac{e^{iks + i\frac{\pi}{4}}}{2\sqrt{2\pi ks}}\left[\sec\tfrac{1}{2}(\theta-\alpha) + \csc\tfrac{1}{2}(\theta+\alpha)\right]\left[(1 - a^{-1}s \sin\theta)\right]^{-1/2}.$$

The result (31) for $u_{\text{mult diff}}$ now applies with F given by (48). As we are considering only normal incidence, $\alpha = 0$ and (31) simplifies to

$$(49)\qquad u_{\text{mult diff}} = \frac{F(2a,0,\pi/2)}{1 - F(2a,\pi/2,\pi/2)}\left[F(r_1,\pi/2,\theta_1) + F(r_2,\pi/2,\theta_2)\right].$$

At points far from the aperture, (49) becomes

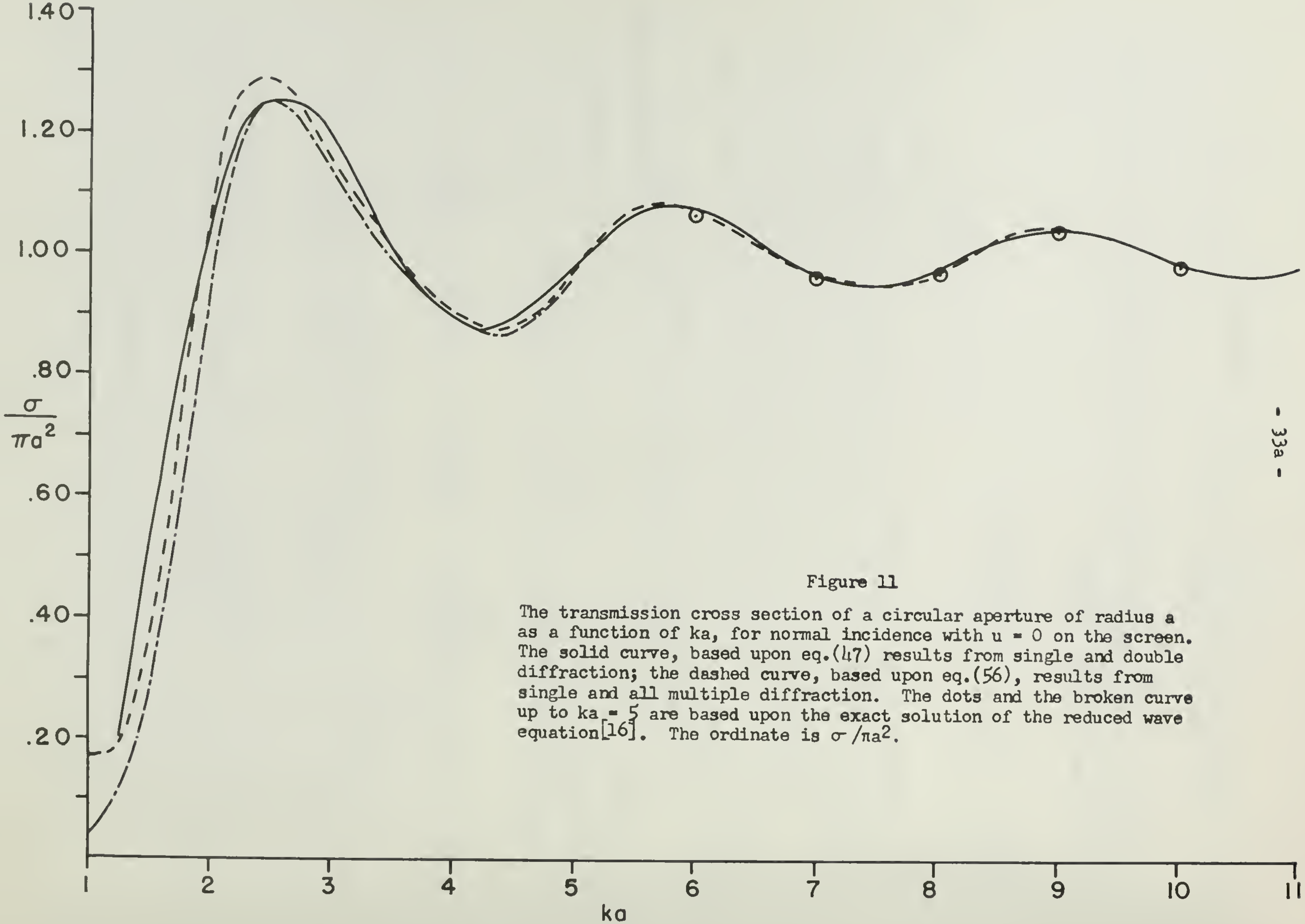


Figure 11

The transmission cross section of a circular aperture of radius a as a function of ka, for normal incidence with u = 0 on the screen. The solid curve, based upon eq.(47) results from single and double diffraction; the dashed curve, based upon eq.(56), results from single and all multiple diffraction. The dots and the broken curve up to ka = 5 are based upon the exact solution of the reduced wave equation[16]. The ordinate is $\sigma/\pi a^2$.

$$u_{\text{mult diff}} = - \frac{ke^{ikr}}{2\pi r} f_m(\phi). \tag{50}$$

The diffraction pattern $f_m(\phi)$ due to multiple diffraction is given by

$$f_m(\phi) = -4 \sqrt{2\pi a}\, k^{-3/2} \left[2 \sqrt{\pi ka}\, e^{-i2ka + i\pi/4} + 1\right]^{-1} \frac{\cos \frac{\phi}{2} \cos(ka \sin \phi - \frac{\pi}{4}) - i \sin \frac{\phi}{2} \sin(ka \sin \phi - \frac{\pi}{4})}{\cos \phi \sqrt{\sin \phi}}. \tag{51}$$

The total diffraction pattern $f(\phi) = f_s(\phi) + f_m(\phi)$. From (51) and (38) we find

$$|f(\phi)| = \sqrt{2\pi a}\, k^{-3/2} (\sin \phi)^{-1/2} \left\{\left[\frac{\cos \eta}{\cos \frac{\phi}{2}} - \frac{2\mu \cos \frac{\phi}{2} \cos \eta + 2\nu \sin \frac{\phi}{2} \sin \eta}{\cos \phi}\right]^2 + \left[\frac{\sin \eta}{\sin \frac{\phi}{2}} - \frac{2\nu \cos \frac{\phi}{2} \cos \eta - 2\mu \sin \frac{\phi}{2} \sin \eta}{\cos \phi}\right]^2\right\}^{1/2} \tag{52}$$

In (52) we have used the abbreviations $\eta = ka \sin \phi - \frac{\pi}{4}$ and

$$\mu = \frac{4 \sqrt{\pi ka} \cos(2ka - \frac{\pi}{4}) + 2}{4\pi ka + 4 \sqrt{\pi ka} \cos(2ka - \frac{\pi}{4}) + 1} \tag{53}$$

$$\nu = \frac{4 \sqrt{\pi ka} \sin(2ka - \frac{\pi}{4})}{4\pi ka + 4 \sqrt{\pi ka} \cos(2ka - \frac{\pi}{4}) + 1}$$

It is to be noted that $f_m(\phi)$ becomes infinite on the axis $\phi = 0$, which is a caustic of the multiply diffracted rays, and also on the screen $\phi = \frac{\pi}{2}$, which is like a shadow boundary. To compute the transmission cross section by the cross section theorem we must correct $f_m(\phi)$ at $\phi = 0$. To this end we again observe that near the axis $u_{\text{mult diff}}$ becomes infinite just like $u_{\text{single diff}}$ and $u_{\text{double diff}}$. In fact by comparing (49) and (44) we see that

$$(54) \quad u_{\text{mult diff}} = \left[1 - F(2a,\pi/2,\pi/2)\right]^{-1} u_{\text{double diff}}.$$

Thus we can obtain $u_{\text{mult diff}}$ near the axis and far from the screen by using (46) in (54). This yields the following value for $f_m(\phi)$ near $\phi = 0$:

$$(55) \quad f_m(\phi) = -\frac{4\pi a}{k}\left[2\sqrt{\pi ka}\, e^{-i2ka + i\pi/4} + 1\right]^{-1} J_o(ka \sin \phi).$$

Now applying the cross section theorem to (55) and adding the contribution πa^2 from (42), we obtain

$$(56) \quad \frac{\sigma}{\pi a^2} = 1 - \frac{2 \sin (2ka - \frac{\pi}{4})}{\sqrt{\pi}\,(ka)^{3/2}}\left[1 + (\pi ka)^{-1/2} \cos(2ka - \frac{\pi}{4}) + (4\pi ka)^{-1}\right]^{-1}.$$

The cross section given by (56) is shown as a function of ka in Figure 11, along with the result (47) due only to single and double diffraction.

We see from (49) that $u_{\text{mult diff}}$ is 'infinite' if the factor in the denominator becomes zero. Just as in the case of the slit, this determines eigenvalues of k which are complex. The equation they satisfy is $2\sqrt{\pi ka}\, e^{-i2ka + i\pi/4} + 1 = 0$. None of the roots of this equation has a positive imaginary part, so they all yield modes which decay in time.

9. Diffraction by a convex aperture

Suppose that the rim of the aperture is a smooth, closed, convex curve. Then for a normally incident plane wave it is easy to find the singly diffracted rays through any point P. There are just two such rays, and they come from the two points on the rim which are nearest to and farthest from P. Thus the singly diffracted field contains two terms, so it is of the form (36). However the radius a in each term of (36) now denotes the radius of curvature ρ of the rim at one of the points of diffraction. In case the rim has a center of symmetry, as, for example an ellipse does, these two radii are equal. The field far from the aperture can then be simplified to

$$(54)\qquad u_{\text{single diff}}(P) = -\frac{e^{ikr}}{r}\sqrt{\frac{\rho(\Psi)}{2\pi k\sin\phi}}\left[i\,\frac{\sin\left[ka(\Psi)\sin\phi-\frac{\pi}{4}\right]}{\sin\frac{\phi}{2}}\pm\frac{\cos\left[ka(\Psi)\sin\phi-\frac{\pi}{4}\right]}{\cos\frac{\phi}{2}}\right]$$

In equation (54), which applies to a scalar field satisfying the reduced wave equation, the upper sign refers to a screen on which $u = 0$, while the lower sign refers to one on which $u_x = 0$. The origin of the polar coordinates r, ϕ and Ψ of P is the center of the aperture. The angle ϕ is the angle between the x-axis and the radius vector from the origin to P. The angle Ψ is the angle between the y-axis and the projection of this radius vector onto the plane of the screen. The radius of curvature of the aperture edge is denoted by $\rho(\Psi)$ and the width of the aperture in the Ψ direction is denoted by $2a(\Psi)$. By the width in the direction Ψ we mean the distance between the two tangents to the edge which are perpendicular to the Ψ direction, and $\rho(\Psi)$ is the radius of curvature at either point of tangency.

Consider now an elliptic aperture with a major axis of length 2b parallel to the y-axis and a minor axis of length $2\,\varepsilon\, b$. Such an aperture has a center of symmetry so (54) applies to it. The curvature and width of this aperture are easily found to be

$$\rho(\Psi) = \varepsilon^{-1} b \left[1+(\varepsilon^2-1)(1+\varepsilon^2\tan^2\Psi)^{-1}\right]^{3/2} \tag{55}$$

$$a(\Psi) = \varepsilon b \left[1+(\varepsilon^2-1)(1+\varepsilon^2\tan^2\Psi)^{-1}\right]^{-1/2} \tag{56}$$

When these expressions are used in (54) an explicit formula for the singly diffracted far field of an elliptic aperture is obtained.

By calculations similar to those made in the cases of the slit and the circular aperture, we could determine the multiply diffracted field for any aperture. However each order of diffraction would require a separate and lengthy calculation. Therefore we will not attempt this, but instead we will consider only the doubly diffracted field produced by an incident plane wave. We will be primarily interested in this field far from the aperture in the forward direction, i.e. in the direction of the incident rays. The reason for this interest is that the far field in the forward direction determines the transmission cross section of the aperture, according to the cross section theorem. Furthermore, the geometrical optics field and the singly diffracted field just yield the geometrical cross section of the aperture, i.e. the projected area of the aperture [8]. Therefore the doubly diffracted field in the forward direction will yield the first correction to the geometrical cross section.

It is interesting to observe that in general only a few doubly diffracted rays go in the forward direction. To show this, let us denote by P a point on the edge at which an incident ray is diffracted and by Q the corresponding point of secondary diffraction. Then one of the diffracted rays produced at P must traverse the segment PQ and be diffracted again at Q. Now a doubly diffracted ray at Q which goes in the forward direction is parallel to the incident ray at P. This is possible, according to the law of diffraction, only if the segment PQ makes the same angle with the edge at Q as it does at P. Furthermore the tangent to the edge at Q must be parallel to the tangent at P. Thus only those singly diffracted rays which satisfy this condition will produce doubly diffracted rays going in the forward direction.

To illustrate this result, let us consider normal incidence. In this case the singly diffracted ray PQ must be normal to the edge at both P and Q. For an ellipse there are just two pairs of points - the ends of the principal axes - having this property. Thus only four doubly diffracted rays will go in the forward direction when a plane wave is normally incident on an elliptic aperture. It is easily seen that there are always at least two such pairs of points whenever the edge is a smooth closed convex curve. One pair lies at the ends of a diameter of maximum width and the other pair at the ends of a diameter of minimum width. Thus for an perture with any such curve as rim, a normally incident plane wave will produce at least four doubly diffracted rays going in the forward direction. In general there will be only four such rays, although in some cases there may be more. Thus for example any aperture of constant width will yield infinitely many - one from each point on the edge. We have already considered the circular aperture, which showed this behavior. On the other hand, for non-normal incidence the circle yields only two doubly diffracted rays in the forward direction.

The field on a ray diffracted at P and then again at Q can be computed by applying (12) twice. The result is, for an incident plane wave of unit amplitude

$$(57) \quad u_{\text{double diff}} = D_Q D_P \left[d(1 + \rho_{1P}^{-1} d)\right]^{-1/2} \left[s(1 + \rho_{1Q}^{-1} s)\right]^{-1/2} e^{ik[\Psi(P) + d + s]}$$

In (57) s denotes distance from Q, d is the distance from P to Q, ρ_{1P} and ρ_{1Q} are the values of ρ_1 at P and Q respectively, and D_P and D_Q are the corresponding values of D. The angle θ at P is $\frac{\pi}{2}$ since the diffracted ray lies in the plane of the screen and therefore at Q, $\alpha = \frac{\pi}{2}$. If the doubly diffracted ray is parallel to the incident ray then β at Q equals β at P and θ at Q equals $\pi + \alpha$ at P. Now using (13) for D with the upper sign, we have when s is large and ρ_{1Q} is finite,

$$(58) \qquad u_{\text{double diff}} = \frac{e^{ik[\Psi(P) + d + s] + i\frac{\pi}{2}(1-n)}}{\pi ks \sin^2\beta \cos\alpha} \sqrt{\left|\frac{\rho_{1Q}\,\rho_{1P}}{d(\rho_{1P} + d)}\right|}$$

Here n is 0,1 or 2 according as none, one or both of $1+d\rho_{1P}^{-1}$ and ρ_{1Q} are negative. It is the number of times the ray goes through a caustic. The angles β and α are evaluated at P.

We can easily compute ρ_{1P} by using (11) and noting that $\cos\beta = I \cdot t$ where I is a unit vector along the incident ray at the edge and t is the unit vector tangent to the edge. Differentiating this relation with respect to arclength along the edge yields $-\dot{\beta}\sin\beta = I \cdot \dot{t}$, since I is constant for an incident plane wave. Now $\dot{t} = \rho^{-1} n$ and $I \cdot n = -\sin\alpha\sin\beta$. Thus $\rho\dot{\beta}\sin\beta = \sin\alpha\sin\beta$. Furthermore δ, which is the angle between the diffracted ray and the normal n, is just $\frac{\pi}{2} - \beta$. Therefore (11) becomes

$$(59) \qquad \rho_{1P} = -\frac{\rho_P \sin\beta}{1 + \sin\alpha}$$

In Appendix V it is shown that

$$(60) \qquad \rho_{1Q} = -\frac{\rho_Q \sin\beta\left[\rho_P \sin\beta - d(1 + \sin\alpha)\right]}{(1 + \sin\alpha)\left[(\rho_P + \rho_Q)\sin\beta - d(1 + \sin\alpha)\right]}$$

From (59) and (60), or from (A20), we can see that at least one of the factors $1+d\rho_{1P}^{-1}$ and ρ_{1Q} is negative, so n = 1 or 2. Upon inserting (59) and (60) into (58), and summing over all doubly diffracted rays in the forward direction, we obtain

(61) $$u_{\text{double diff}} = \sum \frac{e^{ik[\Psi(P)+d+s]+i\frac{\pi}{2}(1-n)}}{\pi ks \cos\alpha \sin\beta} \sqrt{\frac{\rho_P\rho_Q}{d|B|}} .$$

Here B is defined by

(62) $$B = (1 + \sin\alpha)\left[(\rho_P + \rho_Q)\sin\beta - d(1 + \sin\alpha)\right] .$$

From (61) we can obtain the contribution of the doubly diffracted rays to the cross section by using the cross section theorem. We must first remove the spherical wave factor $-ke^{ik(r+\Psi_0)}/2\pi r$ from (61) and then take the imaginary part of the result. In the spherical wave factor Ψ_0 denotes the phase of the incident wave at the origin. Proceeding in this way and adding in the geometrical cross section, we obtain

(63) $$\sigma = \sigma_{\text{geom}} - \frac{2}{k^2} \sum \frac{1}{\cos\alpha \sin\beta} \sin\left\{k[\Psi(P) - \Psi_0+d+s-r]+\varepsilon\right\}\sqrt{\frac{\rho_P\rho_Q}{d|B|}}$$

In (63) we have introduced $\varepsilon = (1-n)\frac{\pi}{2}$ which is $-\frac{\pi}{2}$ if $B < 0$ and as 0 if $B > 0$, since $n = 2$ if $B < 0$ and $n = 1$ if $B > 0$.

Let us apply the result (63) to a circular aperture. In this case only points at opposite ends of a diameter have parallel tangents, and a diffracted ray traverses a diameter only if $\beta = \frac{\pi}{2}$. It follows that only two doubly diffracted rays reach any point, and they come from the ends of the diameter in the plane containing the point, the center and an incident ray. Then if a denotes the radius of the circle we have $\rho_P = \rho_Q = a$, $d = 2a$, $\Psi(P) - \Psi_0 = a \sin\alpha$, $s-r = a\sin\alpha$ and $B = -2a \sin\alpha\ (1+\sin\alpha)$. If α denotes the positive angle between an incident ray and the normal to the screen, then for one of the doubly diffracted rays we must replace the α in these formulas by $-\alpha$. Inserting these expressions into (63) yields for the circle

$$(64)\quad \sigma = \sigma_{\text{geom}} - \frac{1}{k^2 \cos\alpha \sqrt{\sin\alpha}} \left[\frac{\sin[2ka(1-\sin\alpha)]}{\sqrt{1-\sin\alpha}} - \frac{\cos[2ka(1+\sin\alpha)]}{\sqrt{1+\sin\alpha}} \right]$$

It is to be noted that (64) is not valid for normal incidence, since sin α in the denominator is then zero. In that case our previous result (47) applies.

For normal incidence the result (63) simplifies considerably since then $\alpha = 0$, $\beta = \frac{\pi}{2}$, $\Psi(P) = \Psi_o$, $r = s$ and $B = \rho_P + \rho_Q - d$. With these values inserted, (63) becomes

$$(65)\quad \sigma = \sigma_{\text{geom}} - \frac{2}{k^2} \sum \sqrt{\frac{\rho_P \rho_Q}{d\,|\rho_P + \rho_Q - d|}}\, \sin(kd+\varepsilon)$$

Each term in (65) appears twice, once with incidence at P and the other time with incidence at Q. If the aperture has a center of symmetry, then $\rho_P = \rho_Q$.

If the aperture is an ellipse there are four terms in (65). One pair of terms corresponds to the ends of each principal axis. Thus if the length of the major and minor axes are 2a and 2b respectively, and if $e = \left[1 - (\frac{b}{a})^2\right]^{1/2}$ denotes the eccentricity, then (65) yields

$$(66)\quad \sigma = \sigma_{\text{geom}} - \frac{2}{k^2} \left[\frac{1}{e\sqrt{1-e^2}} \sin 2kb + \frac{e^2-1}{e} \cos 2ka \right]$$

The field on each doubly diffracted ray in (61) is proportional to k^{-1} because each edge diffraction introduces a factor $k^{-1/2}$. Therefore the correction to the cross section (63) is proportional to k^{-2} because another factor of k^{-1} is introduced in defining the diffraction pattern. However, if the caustic of the doubly diffracted rays extends in the forward direction (i.e. if ρ_{1Q} is infinite) then (58) is not valid. In such a case the field in that direction is increased by a factor of a power of k which lies between k^0 and $k^{1/2}$. Therefore in these cases the cross section correction is proportional to a power of k between k^{-2} and $k^{-3/2}$. The maximum power, $k^{-3/2}$, occurs in the result for a circular aperture with normal incidence. The same power occurs in the result for a slit, but for a different reason. In the case of a slit, which is two dimensional, the factor $k^{-1/2}$, instead of k^{-1}, is introduced in the definition of the diffraction pattern. This, with the k^{-1} due to diffraction, accounts for the $k^{-3/2}$ in the cross section correction for a slit.

Let us now consider screens on which the normal derivative of the field vanishes. In Sections 7 and 8 we found that for such screens the field was zero on the doubly and multiply diffracted rays. That result is not correct. It was obtained because the assumption that the diffracted field is proportional to the incident field is not valid for such screens when the incident ray lies in the plane of the screen. The correct assumption in this case is that the diffracted field is proportional to the normal derivative of the incident field at the point of diffraction. Therefore instead of (12) we should have for such rays and screens

$$(12') \qquad u_{\text{edge diff}} = D' \frac{\partial A}{\partial x} \left[s(1 + \rho_1^{-1} s)\right]^{-1/2} e^{ik(s+\Psi)}$$

The new diffraction coefficient D' has been introduced in (12'). It may depend upon the nature of the field, the directions of the incident and diffracted rays, and upon k. From dimensional considerations, D' must be proportional to $k^{-3/2}$. It follows that for the slit and the circle, the contribution to the transmission cross section from doubly diffracted rays will be proportional to $k^{-5/2}$. For a general aperture it will be proportional to k^{-3}. The diffraction coefficient D' will be deduced for a scalar field satisfying the reduced wave equation in Part III of this article[12] and applied to diffraction by a slit.

10. Electromagnetic diffraction by a convex aperture

Let us now consider a linearly polarized plane electromagnetic wave normally incident upon a perfectly conducting screen containing a convex aperture. As we have already seen, two rays pass through each point P. Therefore the singly diffracted electric field $E^s(P)$ is the sum of two terms of the type given in equation (12). The diffraction coefficient D is given by equation (A13). In (A13) we must set $\alpha = 0$ because the wave is normally incident. Now if the incident amplitude is unity and if the incident phase is zero at the screen, then $E^s(P)$ is given by

$$(67) \qquad E^s(P) = - \frac{e^{ikr_1 + i\frac{\pi}{4}}}{2\sqrt{2\pi k}} \left[r_1 (1 - \rho_1^{-1} r_1 \sin\theta_1) \right]^{-1/2}$$

$$\begin{pmatrix} \sec\frac{\theta_1}{2} + \csc\frac{\theta_1}{2} & 0 & 0 \\ 0 & \cos\theta_1 \left[\sec\frac{\theta_1}{2} - \csc\frac{\theta_1}{2} \right] & 0 \\ 0 & \sin\theta_1 \left[\sec\frac{\theta_1}{2} - \csc\frac{\theta_1}{2} \right] & 0 \end{pmatrix} \begin{pmatrix} -\sin\Psi \\ -\cos\Psi \\ 0 \end{pmatrix} + \cdot$$

In (67) the term written explicitly refers to one ray through P and the dot indicates a similar term for the other ray. The angle Ψ is the angle between the negative incident electric field and the normal to the edge and ρ_1 denotes the radius of curvature of the edge, both at the point of diffraction. The components of $E^s(P)$ are those in the directions of the tangent t, the normal n and the binormal b at the point of diffraction. The components in the other term are in terms of a basis referred to the other point of diffraction.

Suppose that the aperture has a center of symmetry. Also, let the point P be far from the aperture. Then for such a point P the two points of diffraction are practically symmetric. Therefore the radii of curvature are equal at these two points while t_2 and n_2 are the negatives of t_1 and n_1 respectively. In addition $\theta_1 \approx \pi + \phi$, $\theta_2 \approx \pi - \phi$, $r_1 \approx r - a(\Psi)\sin\phi$ and $r_2 \approx r + a(\Psi)\sin\phi$. Here ϕ is the angle between the normal to the screen and a ray from the center of symmetry to P. The distance between the two tangents to the edge at the points of diffraction - the width of the aperture - is $2a(\Psi)$. When these simplifications are made in (67) it becomes

$$(68)\quad E^s(P) = -\frac{e^{ikr}}{r}\sqrt{\frac{\rho(\Psi)}{2\pi k\sin\phi}}\begin{bmatrix} \left[i\dfrac{\sin\eta}{\sin\frac{\phi}{2}} + \dfrac{\cos\eta}{\sin\frac{\phi}{2}}\right] & 0 & 0 \\ 0 & -\cos\phi\left[i\dfrac{\sin\eta}{\sin\frac{\phi}{2}} - \dfrac{\cos\eta}{\cos\frac{\phi}{2}}\right] & 0 \\ 0 & -\sin\phi\left[i\dfrac{\sin\eta}{\sin\frac{\phi}{2}} - \dfrac{\cos\eta}{\cos\frac{\phi}{2}}\right] & 0 \end{bmatrix}\begin{bmatrix} \sin\Psi \\ \cos\Psi \\ 0 \end{bmatrix}$$

$$= -\frac{e^{ikr}}{r}\sqrt{\frac{\rho(\Psi)}{2\pi k\sin\phi}}\begin{bmatrix} \left[i\dfrac{\sin\eta}{\sin\frac{\phi}{2}} + \dfrac{\cos\eta}{\cos\frac{\phi}{2}}\right]\sin\Psi \\ -\cos\phi\left[i\dfrac{\sin\eta}{\sin\frac{\phi}{2}} - \dfrac{\cos\eta}{\cos\frac{\phi}{2}}\right]\cos\Psi \\ -\sin\phi\left[i\dfrac{\sin\eta}{\sin\frac{\phi}{2}} - \dfrac{\cos\eta}{\cos\frac{\phi}{2}}\right]\cos\Psi \end{bmatrix}$$

Here $\eta = ka(\Psi)\sin\phi - \frac{\pi}{4}$.

In (68) the components of $E^s(P)$ are given in terms of the basis t, n, b at either point of diffraction. This basis may just as well be located at the center. Its orientation is determined by the condition that Ψ is the angle between the negative incident electric field and n. Furthermore, Ψ is just the angle between the incident electric field and the projection on the screen of the ray from the center to P. This is so because this projection is practically parallel to the normal at the point of diffraction. Therefore if the y axis is parallel to the incident electric field, then r, ϕ and Ψ are just the polar coordinates of P.

From (68) we see that in the H-plane, $\Psi = \pi/2$, the electric field has just one non-zero component which is normal to this plane. Quantitatively this component is exactly the same as the scalar field diffracted by an aperture in a screen on which $u = 0$ (see equation (54)). Qualitatively it is somewhat similar to the field diffracted by a slit of width $2a(\pi/2)$ when the incident electric field is parallel to the edges. On the other hand in the E-plane, $\Psi = 0$, the magnetic field has just one non-zero component, which is also normal to this plane. Quantitatively it is the same as the scalar field diffracted by a hole in a screen on which $u_x = 0$ (see (54)). Qualitatively it is also similar to the field diffracted by a slit of width $2a(0)$ when the incident magnetic field is parallel to the edges. These conclusions follow at once from a consideration of the location of the diffracted rays and the orientation of the incident field relative to the edge at each point of diffraction. In particular, for a circular hole, they apply everywhere and not merely in the far field because at every point P the two points of diffraction are symmetric about the center.

It is convenient to express $E_s(P)$ in terms of the unit vectors $\hat{r}$, $\hat{\phi}$ and $\hat{\Psi}$ of the spherical coordinate system. The transformation equations for a point P in the n,b plane are $t = \hat{\Psi}$, $n = -\hat{r} \sin \phi - \hat{\phi} \cos \phi$ and $b = \hat{r} \cos \phi - \hat{\phi} \sin \phi$. By using these relations in (68) we obtain

$$E^s_\Psi(P) = -\frac{e^{ikr}}{r}\sqrt{\frac{\rho(\Psi)}{2\pi k \sin\phi}}\left(i\,\frac{\sin\eta}{\sin\frac{\phi}{2}} + \frac{\cos\eta}{\cos\frac{\phi}{2}}\right)\sin\Psi \tag{69}$$

$$E^s_\phi(P) = -\frac{e^{ikr}}{r}\sqrt{\frac{\rho(\Psi)}{2\pi k \sin\phi}}\left(i\,\frac{\sin\eta}{\sin\frac{\phi}{2}} - \frac{\cos\eta}{\cos\frac{\phi}{2}}\right)\cos\Psi$$

$$E^s_r(P) = 0$$

These equations show clearly that the diffracted field is transverse to the $\hat{r}$ direction, which is the direction of the diffracted rays far from the aperture. Furthermore, they show that the components of the incident field in and perpendicular to the $\hat{\Psi}$ direction give rise, respectively, to the $\hat{\Psi}$ and $\hat{\phi}$ components of E^s.

11. Diffraction by a grating of slits

Suppose the screen contains a number of parallel slits, all of the same width, 2a, and uniformly spaced. Such an array of slits is a diffraction grating. We can apply our results to determine the field diffracted by such a grating. In particular we can find the positions and intensities of the maxima in the diffraction pattern. These maxima are the spectral orders produced by the grating.

Let us begin by assuming that the incident rays lie in planes normal to the edges of the slits. Then the problem is two-dimensional, so we will restrict our attention to the xy-plane. Suppose that there are 2N+1 slits and that the centers of adjacent slits are separated by the distance b. If the center of the middle slit is placed at $y = 0$, and if the slits are numbered from $t = -N$ to $t = N$, then the center of slit t is at tb. It is convenient to introduce the polar coordinates r_t and ϕ_t with the origin at the center of slit t and with the polar axis parallel to the x-axis.

The singly diffracted field is the sum of the fields on all the rays singly diffracted from any of the slit edges. To compute it we first add together the fields on the two rays from the two edges of each slit. This gives the singly diffracted field of each slit. Then we add together the singly diffracted fields of the various slits. If we consider the field far from the grating, then it suffices to add together the far fields of the various slits. For a plane wave incident at angle α, the far field of slit t is given by equation (22) with r and ϕ replaced by r_t and ϕ_t, and with the extra factor $e^{-ik\,tb\,\sin\alpha}$ to account for the phase of the incident wave. Thus we have

$$u_{\text{single diff}} = -\sum_{t=-N}^{t=N} \sqrt{\frac{k}{2\pi r_t}}\, e^{ik(r_t - tb\sin\alpha) + i\frac{\pi}{4}} f_s(\phi_t). \tag{70}$$

In order to simplify (70) we make use of the polar coordinates r and ϕ with their origin at the center of the middle slit. At points far from the grating compared to the grating length, which is $2Nb + 2a$, the angles ϕ_t are all nearly equal to ϕ. Furthermore r_t is approximately $r - tb \sin\phi$. When these substitutions are made in (70) we have instead

$$\begin{aligned} u_{\text{single diff}} &= -\sqrt{\frac{k}{2\pi r}}\, e^{ikr+i\frac{\pi}{4}} f_s(\phi) \sum_{t=-N}^{t=N} e^{-iktb(\sin\phi+\sin\alpha)} \\ &= -\sqrt{\frac{k}{2\pi r}}\, e^{ikr+i\frac{\pi}{4}} f_s(\phi)\, \frac{\sin\left[(N+\frac{1}{2})kb(\sin\phi+\sin\alpha)\right]}{\sin\left[\frac{1}{2}kb(\sin\phi+\sin\alpha)\right]} . \end{aligned} \tag{71}$$

The same result applies when the grating has an even number of slits, say 2N, provided that the origin is at the midpoint of the grating and that $N + \frac{1}{2}$ is replaced by N.

The result (71) shows that the diffraction pattern of the grating is the product of the diffraction pattern of a single slit and the grating factor, which is the quotient of sines. It is clear that the same result with $f_s(\phi)$ replaced by $f_m(\phi)$ would be obtained if we took account of rays multiply diffracted between the two edges of each slit. However a different result would be obtained if we also included rays multiply diffracted between edges of different slits.

12. Conclusion

The method which we have applied in this article to diffraction by an aperture is a natural extension of geometrical optics. It augments geometrical optics by introducing diffracted rays and by assigning a field to each ray. Both of these aspects can be worked out by the use of elementary mathematics alone. Nevertheless it leads directly to results which have only recently been obtained by elaborate mathematical methods, and to others which have not yet been obtained in any other way.

In Part II the various " physical optics " theories based upon the Kirchhoff method will be examined. Those theories require the evaluation of integrals. It will be shown that when those integrals are evaluated for small wavelengths they lead to results which have an immediate interpretation in terms of our singly diffracted rays. However the quantitative value of the field obtained from the integrals will be incorrect, as is well known. It therefore seems that these integral methods have certain disadvantages which our method by-passes. They are harder to use, yield quantitatively inferior results and cannot take account of multiple diffraction. On the other hand, they do yield finite - but incorrect - field values on caustics and shadow boundaries.

Mathematically we can interpret the field constructed by our method from the viewpoint of asymptotic expansions. Let us suppose that the field u is governed by the reduced wave equation $(\triangle + k^2)u = 0$. The field which we have constructed consists of a sum of terms, one associated with each family of rays. Each term satisfies the reduced wave equation up to terms in k^{-1}, as has been shown elsewhere[4]. It is to be expected that a series of additional terms proportional to higher powers of k^{-1} should be added to the amplitude of each of the terms we have found. This has been done in [4] for other kinds of diffraction problems. The infinite series of terms associated with each family of rays should be so chosen as to satisfy the reduced wave equation formally, but need not necessarily converge. The sum of the various series corresponding to the various families of rays should then constitute an asymptotic expansion of the solution of the reduced wave equation.

The fact that our expressions become infinite on certain surfaces can then be interpreted as being due to a change of asymptotic behavior of the solution. These surfaces may therefore be called " Stokes surfaces" , in analogy with the terminology in the theory of ordinary differential equations. Shadow boundaries are examples of such surfaces. Certain caustics are also of this type. Others correspond to a different asymptotic behavior of the solution on the caustic itself, the fields on the two sides of it having the same form.

An interesting feature of our theory is its relation to what we have called canonical problems. These are special diffraction problems with relatively simple incident fields and boundary shapes. By means of our method, certain coefficients obtained from these special problems can be used in the analysis of many other more complicated problems. Thus a new motivation is provided for solving these special problems, since their solutions now have a more general significance.

Appendix I. Caustics of diffracted rays

We wish to consider an edge given by the vector $x(s)$, from each point of which a cone of diffracted rays emerges. Let $\beta(s)$ denote the half-angle of the cone with its vertex at $x(s)$. The axis of this cone is the tangent $\dot{x}(s)$. If y denotes a variable point on this cone, then y satisfies the equation

$$\text{(A1)} \qquad \frac{y - x(s)}{|y - x(s)|} \cdot \dot{x}(s) = \cos \beta(s)$$

The envelope or caustic of the diffracted rays is also the envelope of these cones, since these cones contain the diffracted rays. To find this envelope we differentiate (A1) with respect to s and obtain

$$\text{(A2)} \qquad (y - x) \cdot \ddot{x} - \dot{x}^2 = (\cos \beta)^{\dot{}} \; |y - x| - \frac{y - x}{|y - x|} \cdot \dot{x} \cos \beta$$

Now using (A1) in (A2) yields

$$\text{(A3)} \qquad (y - x) \cdot (\ddot{x} - \dot{x}[\log \cos \beta]^{\dot{}}) = \dot{x}^2 - \cos^2 \beta$$

If the parameter s denotes arclength along the edge then $\dot{x}^2 = 1$, $\ddot{x} = \rho^{-1} n$ and $\dot{x} = t$. Here ρ, n and t are the radius of curvature, the unit normal and the unit tangent to the edge, respectively. Introducing this notation into (A1) and (A3) we obtain as the equations of the caustic

$$\text{(A4)} \qquad (y - x) \cdot [n - t\rho(\log \cos \beta)^{\dot{}}] = \rho \sin^2 \beta$$

$$\text{(A5)} \qquad (y - x) \cdot t = |y - x| \cos \beta$$

From (A4) and (A5) we can determine the distance ρ_1 from the edge to the caustic along any particular ray. This distance ρ_1 is just $-|y - x|$ if we take account of the direction in which ρ_1 is measured positively. Thus we obtain

$$\text{(A6)} \qquad \rho_1 = - \frac{\rho \sin^2\beta}{\rho\dot{\beta} \sin \beta + \cos \delta}$$

The angle δ introduced in (A6) is defined by the equation

$$\text{(A7)} \qquad \cos \delta = \frac{(y - x)}{|y - x|} \cdot n$$

It is just the angle between the diffracted ray and the normal to the edge.

As an example of the use of (A4) and (A5), let us suppose that the edge is a plane curve, and that a plane wave is normally incident. Then $\beta(s) = \frac{\pi}{2}$ so $\dot{\beta} = 0$. Equations (A4) and (A5) then become $(y - x)\cdot n = \rho$ and $(y - x)\cdot t = 0$. The first of these equations shows that the caustic contains the curve of centers of curvature of the edge, i.e., the evolute of the edge. The second equation shows that the caustic consists of the set of straight lines through these points, each normal to the plane of the edge. Thus the caustic is a cylinder with generators normal to the plane of the edge, the cross section of which is the evolute of the edge. If the edge is the edge of a thin flat obstacle, then the caustic will appear as a bright line in any cross section of its shadow. Bright lines of exactly this shape have been observed by J. Coulson and G. G. Becknell[13]. The special case of a circular disc is well known. The evolute in this case is just the center and the caustic is the axis, which appears as a bright spot in a cross section of the shadow.

In the case of oblique incidence or of a non-planar edge, it may be shown from (A4) and A5) that with increasing distance behind the obstacle the caustic approaches a cylinder. The generators of this cylinder are parallel to the incident rays and its cross section is the evolute of the shadow cross section. Thus in any normal cross section of the shadow, far enough from the obstacle, a bright line should be seen. This bright line should coincide with the evolute of the shadow boundary. Such bright lines were also observed by G. G. Becknell and J. Coulson[14]. However, they did not observe the departure from this shape which must occur nearer to the obstacle, as all their observations were made far behind it.

Appendix II. The Diffraction Coefficient D

It has been noted that the diffraction coefficient D depends upon the nature of the field and the kind of screen. Let us therefore consider only scalar fields governed by the reduced wave equation, and screens on which either u or its normal derivative vanishes. For such screens and fields the two dimensional problem of diffraction of a plane wave by a screen in the form of a half plane has been solved exactly. By a two dimensional problem we mean one in which each incident ray lies in a plane normal to the edge. By solved we mean that a solution of the reduced wave equation satisfying the boundary and other conditions has been constructed.

The solution for either boundary condition can be extended to apply to the three dimensional case in which the incident rays lie in planes making the angle β with the edge. To this end it is merely necessary to replace k in the solution by $k \sin \beta$ and then to multiply the result by $e^{ikz \cos \beta}$, provided that the z-axis is parallel to the edge.

Once the solution has been obtained it can be expanded asymptotically in k, for k large. If the amplitude of the incident wave is unity the result is

$$\text{(A7)} \qquad u \sim u_{geom} - \frac{e^{ik(z \cos \beta + r \sin \beta) + i\pi/4}}{2\sqrt{2\pi kr} \ \sin \beta} \left[\sec \frac{1}{2} (\theta - \alpha) \pm \csc \frac{1}{2} (\theta + \alpha) \right]$$

Here u_{geom} is the field determined by the incident and reflected rays and r is radial distance from the edge. The angles are shown in Figure 6 and the upper sign applies to the screen on which $u = 0$. This is to be compared with the solution given by the present method. In this case there is just one diffracted ray through each point. Also $\dot{\beta} = 0$ so it follows that $\rho_1^{-1} = 0$. Then using (12) we obtain

$$\text{(A8)} \qquad u = u_{geom} + \frac{De^{ik(z_0 \cos \beta + s)}}{\sqrt{s}}$$

Here z_0 denotes the point of diffraction on the edge. From the law of diffraction we see that $s = (z - z_0)\cos \beta + r \sin \beta = r/\sin \beta$. Upon inserting the first of these expressions for s into the exponent in (A8) and the second expression into the denominator, our result (A8) becomes

$$\text{(A9)} \qquad u \sim u_{geom} + \frac{De^{ik(z \cos \beta + r \sin \beta)}}{\sqrt{r/\sin \beta}}$$

The result (A9) is seen to depend upon z and r in exactly the same way as does (A7) and to agree with it precisely if D has the value given in equation (13).

A similar consideration can be applied to diffraction by a wedge of angle $(2-n)\pi$ instead of a screen, which is a wedge of angle $(2-2)\pi = 0$. In this way one obtains the diffraction coefficient

$$(A10) \qquad D = -\frac{e^{i\pi/4}}{2\sqrt{2\pi k}\,\sin\beta}\left[\left(\cos\frac{\pi}{n} - \cos\frac{\theta-\alpha}{n}\right)^{-1} \mp \left(\cos\frac{\pi}{n} - \cos\frac{\theta+\alpha+\pi}{n}\right)^{-1}\right]$$

Here again the upper sign applies to the screen on which $u = 0$ and the lower sign to the screen on which the normal derivative of u is zero. The angles are defined as in Figure 6.

Appendix III. The electromagnetic diffraction matrix

The electromagnetic problem of diffraction of an arbitrary plane wave by a perfectly conducting half-plane has been solved exactly.[15] This means that an exact solution of Maxwell's equations satisfying the boundary and other condtions of the problem has been constructed for a plane wave travelling in any direction. This solution has also been expanded asymptotically for large k.

This same problem can be treated by the present method. In this case there is exactly one diffracted ray through each point not lying on the edge. Furthermore, $\beta = 0$ and $\rho^{-1} = 0$. We then find that the electric field E is given by

$$(A11) \qquad E(P) = E_{geom}(P) + \frac{e^{iks}}{\sqrt{s}}\, D\, E_{inc}(Q) \quad .$$

In this equation Q denotes the point on the edge at which the diffracted ray through P is produced. The distance from P to Q is s, the geometrical optics field is denoted by $E_{geom}(P)$ and the diffraction matrix is denoted by D.

We now compare the expanded form of the exact solution with our result. We find that the two are identical if D is given by

$$D=\frac{-e^{i\pi/4}}{2\sin\beta\sqrt{2\pi k}}\begin{bmatrix} \sec\frac{1}{2}(\theta-\alpha)+\csc\frac{1}{2}(\theta+\alpha) & 0 & 0 \\ -2\cot\beta\,\dfrac{\cos\frac{\theta}{2}+\sin\frac{\theta}{2}}{\cos\frac{\alpha}{2}+\sin\frac{\alpha}{2}} & \dfrac{\cos\theta}{\cos\alpha}\left[\sec\frac{1}{2}(\theta-\alpha)-\csc\frac{1}{2}(\theta+\alpha)\right] & 0 \\ \cot\beta\left[2\dfrac{\cos\frac{\theta}{2}-\sin\frac{\theta}{2}}{\cos\frac{\alpha}{2}+\sin\frac{\alpha}{2}}-\dfrac{\sec\frac{1}{2}(\theta-\alpha)-\csc\frac{1}{2}(\theta+\alpha)}{\cos\alpha}\right] & \dfrac{\sin\theta}{\cos\alpha}\left[\sec\frac{1}{2}(\theta-\alpha)-\csc\frac{1}{2}(\theta+\alpha)\right] & 0 \end{bmatrix} \tag{A12}$$

This form of D applies if E is expressed in terms of the basis (t,n,b). Here t denotes the unit vector tangent to the edge at the point of diffraction Q, n is the unit normal to the edge (pointing into the aperture) and b is the unit binormal to the edge, i.e., a normal to the screen pointing toward the dark side. The direction of t is determined by the condition that these three vectors form a right-handed system, i.e., $n \times b = t$.

The possibility of defining D in such a way that our result coincides with the expansion of the exact solution is a partial check on our method. It also follows from the explicit form of D that E is transverse to the diffracted ray, as we have assumed. The expression for D given in (A12) is the desired form of the electromagnetic diffraction matrix.

It is to be noted that (A12) simplifies when $\beta = \pi/2$, to the following

$$D=\frac{-e^{i\pi/4}}{2\sqrt{2\pi k}}\begin{bmatrix} \sec\frac{1}{2}(\theta-\alpha)+\csc\frac{1}{2}(\theta+\alpha) & 0 & 0 \\ 0 & \dfrac{\cos\theta}{\cos\alpha}\left[\sec\frac{1}{2}(\theta-\alpha)-\csc\frac{1}{2}(\theta+\alpha)\right] & 0 \\ 0 & \dfrac{\sin\theta}{\cos\alpha}\left[\sec\frac{1}{2}(\theta-\alpha)-\csc\frac{1}{2}(\theta+\alpha)\right] & 0 \end{bmatrix} \tag{A13}$$

This form of D can also be obtained by reducing the electromagnetic problem to two scalar problems, one for E_t and the other for H_t, the components of E and H along the edge. Each of these components satisfies the reduced wave equation, while on a perfect conductor E_t and the normal derivative of H_t are zero. Therefore the scalar diffraction coefficients in equation (13) apply to these two components. If E_n and E_b are computed from H_t and if the scalar diffraction coefficients are used, the result (A13) is again obtained for the electromagnetic diffraction matrix.

Appendix IV. The field on an axial caustic

As has been pointed out before, the field given by the present method becomes infinite on a caustic. A finite value for the field on the caustic, and a corrected value for the field near the caustic, can be obtained by the use of a canonical solution, as was mentioned before. In this method an exact solution of the equation or equations governing the field is required. When this solution is expanded asymptotically, for large k, it must correspond to a system of rays possessing a caustic of the desired type. A comparison of the exact solution with its expansion shows how the expansion must be modified to remain finite on the caustic. It is assumed that the same modification applies to any field with a similar caustic.

Let us apply this method to fields satisfying the reduced wave equation and having an axial caustic. We begin with an exact solution of the reduced wave equation

$$u = e^{ikx \sin \delta} J_n(k\rho \cos \delta) \cos n\phi \quad . \tag{A14}$$

In (A14) J_n denotes the Bessel function of integral order n; x, ρ and ϕ are cylindrical coordinates, and δ is a constant. The asymptotic expansion of (A14) for k large, obtained by inserting the expansion of J_n into (A14) is

$$(A15)\qquad u \sim \frac{\cos n\phi}{\sqrt{2\pi k\rho \cos \delta}} \left[\exp\left[-ik(\rho \cos \delta - x \sin \delta) + i(n + \tfrac{1}{2})\tfrac{\pi}{2}\right] + \right.$$
$$\left. + \exp\left[ik(\rho \cos \delta + x \sin \delta) - i(n + \tfrac{1}{2})\tfrac{\pi}{2}\right]\right]$$
$$= \frac{\cos n\phi}{\sqrt{2\pi k\rho \cos \delta}} \; 2\, e^{ikx \sin \delta} \cos\left[k\rho \cos \delta - (n + \tfrac{1}{2})\tfrac{\pi}{2}\right] \quad .$$

The first form of (A15) shows that u is the field on a set of rays which converge symmetrically on the x-axis, pass through it, and then diverge. The constant δ is the angle between any ray and a normal to the axis, so $\frac{\pi}{2} - \delta$ is the angle between each ray and the axis. The field at each point off the axis is the sum of the fields on two rays, one going toward the axis and the other already diverging. For such a set of rays, the method we have given for constructing the field yields exactly the expressions given in equation (A15). This can be seen from the right side of equation (10) by making ρ_1 infinite. This agreement provides additional support for our method.

The expression for u in (A15) is infinite at $\rho = 0$, while actually u has there the finite value given by (A14). To obtain the finite result in (A14) from the divergent result in (A15), we must multiply the field given in (A15) by

$$(A16)\qquad \frac{1}{2}\sqrt{2\pi k\rho \cos \delta}\, \sec\left[k\rho \cos \delta - (n + \tfrac{1}{2})\tfrac{\pi}{2}\right] J_n(k\rho \cos \delta).$$

We assume that any field governed by the reduced wave equation and having an axial caustic can be corrected on and near the axis by the same factor (A16). The value of n to be used in (A16) is determined by the angular dependence of the field under consideration.

It is to be noted that the phase of u given by (A15) changes by exactly $-\frac{\pi}{2}$ as a ray passes through the caustic. The apparent change of $n\frac{\pi}{2}$ is exactly cancelled by a corresponding change in cos $n\phi$ since ϕ changes by π. This result agrees exactly with our equation (12) which yields a phase change of $-\frac{\pi}{2}$ whenever a ray passes through a caustic. In (12) one of the factors s or $1+\rho_1^{-1}s$ under the radical in the denominator changes sign on passing through a caustic, yielding the phase change. The agreement of (A15) with this result is another check on our theory.

Appendix V. Caustics of doubly diffracted rays

When a plane wave is incident upon an aperture, some of the singly diffracted rays which it produces cross the aperture and hit the edge again producing doubly diffracted rays. Consider a ray which is initially diffracted at a point P on the edge and then diffracted again at another point Q. We will suppose that one doubly diffracted ray leaves Q in the forward direction, i.e. in the direction of the incident ray at P. We wish to compute ρ_{1Q}, the distance to the caustic of the doubly diffracted rays, along the ray which leaves Q in the forward direction. This distance is given by equation (11) in which all quantities are to be evaluated at Q. We have already seen that β at Q is equal to β at P, and that the cosine of the angle between the incident ray and the normal to the edge at P is $-\sin\alpha \sin\beta$. Since the diffracted ray at Q is parallel to the incident ray at P, and since the normal at Q is the negative of the normal at P, it follows that at Q, $\cos\delta = \sin\alpha\sin\beta$.

In order to compute $\dot{\beta}$, we note that the center of curvature on the ray incident at Q is at the distance $d + \rho_{1P}$ from Q. Now $\cos\beta = I\cdot t$ where I is a unit vector along the incident ray and t is the unit tangent to the edge. Thus $-\dot{\beta}\sin\beta = \dot{I}\cdot t + I\cdot\dot{t} = \dot{I}\cdot t + I\cdot n/\rho_Q$. As we found in studying spherical waves, preceding equation (20), $\dot{I}\cdot t = \frac{\sin^2\beta}{\rho_{1P}+d}$. Furthermore, $I\cdot n = -\sin\beta$. Combining these results we have

$$\text{(A17)} \qquad -\dot{\beta} \sin \beta = \frac{\sin^2\beta}{\rho_{1P}+d} - \frac{\sin \beta}{\rho_Q}$$

Inserting (A17) and the value of cos δ obtained above into (11) yields

$$\text{(A18)} \qquad \frac{1}{\rho_{1Q}} = \frac{1}{\rho_{1P}+d} - \frac{1+\sin \alpha}{\rho_Q \sin \beta}$$

If (A18) is solved for ρ_{1Q} and if ρ_{1P} is inserted from (59), equation (60) is obtained.

Equation (A18) is precisely the mirror formula of optics, if we define the focal length f_Q by

$$\text{(A19)} \qquad f_Q = -\frac{\rho_Q \sin \beta}{1+\sin \alpha}$$

Then (A18) becomes

$$\text{(A20)} \qquad \frac{1}{\rho_{1Q}} = \frac{1}{\rho_{1P}+d} + \frac{1}{f_Q}$$

By comparing (A19) with equation (59) we see that f_Q is indeed a focal length. For if a plane wave hits the edge at Q, coming from the direction towards which the diffracted ray goes, then f_Q is the distance to the caustic along the diffracted ray from Q to P. The distance $\rho_{1P}+d$ is clearly the 'object distance,' since it is the distance to the center of curvature of the incident rays. Similarly ρ_{1Q} is the 'image' distance because it is the distance to the caustic.

References

[1] Keller, J.B. - The geometrical theory of diffraction; Symposium on Microwave Optics, Eaton Electronics Research Laboratory, McGill University, Montreal, Canada, June, 1953.

[2] Keller, J.B. - Diffraction by a convex cylinder; URSI Michigan Symposium on Electromagnetic Wave Propagation, Ann Arbor, June, 1955. To be published in Trans. Inst. Radio Engrs., P.G.A.P.

[3] Keller, J.B. and Levy, B. - Diffraction by a smooth object; to be published, New York University, Institute of Mathematical Sciences, Division of Electromagnetic Research, EM Series (1956).

[4] Keller, J.B., Lewis, R.M. and Seckler, B.D. - Asymptotic solution of some diffraction problems; Comm. Pure and Appl. Math, 9, 207 (1956); also New York University, Institute of Mathematical Sciences, Division of Electromagnetic Research, Report No. EM-81 (1955).

[5] Sommerfeld, A.J.W. - Optics; Academic Press, New York, 1954, p. 249.

[6] Kraus, L. - Diffraction by a plane angular sector; Ph.D. thesis, New York University, New York, 1955.

[7] Braunbek, W. - Neue Näherungsmethode für die Beugung am ebenen Schirm; Zeits. für Physik, 127, 381 (1950).

- Zur Beugung an der Kreisscheibe; ibid., 127, 405 (1950).

[8] Keller, J.B., Lewis, R.M. and Seckler, B.D. - Diffraction by an aperture II; to be published, New York University, Institute of Mathematical Sciences, Division of Electromagnetic Research, EM Series (1956).

[9] Levine, H. - Unpublished.

[10] Karp, S.N. and Russek, A. - Diffraction by a wide slit; New York University, Institute of Mathematical Sciences, Division of Electromagnetic Research, Research Report No. EM-75(1955).

[11] Levine, H. - Diffraction by a circular aperture at high frequencies; New York University, Institute of Mathematical Sciences, Division of Electromagnetic Research, Research Report No. EM-84 (1955).

[12] Karp, S.N. and Keller, J.B. - Diffraction by an aperture III; to be published, New York University, Institute of Mathematical Sciences, Division of Electromagnetic Research, EM Series (1956).

[13] Coulson, J. and Becknell, G.G. - Reciprocal diffraction relations between circular and elliptical plates; Phys. Rev., 20, 594 (1922)

[14] Coulson, J. and Becknell, G.G. - An extension of the principle of the diffraction evolute and some of its structural detail; ibid., p. 607.

[15] Jones, D.S. - Note on diffraction by an edge; Quar. Journ. Mech. and Appl. Math., 3, 420 (1950).

[16] Bouwkamp, C.J. - Diffraction theory; Reports on Prog. in Physics, 17, 35-100 (1954).

NYU
EM-
92

c. 1

AUTHOR Keller

TITLE Diffraction by an aperture. I.

DATE DUE | BORROWER'S N

NYU
EM-
92
c. 1

Keller

Diffraction by an aperture I.

Date Due

DEC 18 '56			
[illegible]			
[illegible]			
11/18/66 ILL			
FMC Corp			
[illegible] 1987			
JAN 4 1986			
JAN 16 1986			
FEB 7 1986			
MAR 5 1986			
	PRINTED	IN U. S. A.	